

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
14 February 2002 (14.02.2002)

PCT

(10) International Publication Number  
**WO 02/12342 A2**

(51) International Patent Classification<sup>7</sup>: **C07K 14/705**,  
C12Q 1/68, C12N 15/12, 5/10, C07K 16/28, G06F 17/00

(74) Agents: **SHANER, Sandra, L.** et al.; Genaissance Pharmaceuticals, Inc., Five Science Park, New Haven, CT 06511 (US).

(21) International Application Number: PCT/US01/24649

(22) International Filing Date: 6 August 2001 (06.08.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
60/223,177 4 August 2000 (04.08.2000) US

(71) Applicant (*for all designated States except US*): **GENAISANCE PHARMACEUTICALS, INC.** [US/US]; Five Science Park, New Haven, CT 06511 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published:**

— *without international search report and to be republished upon receipt of that report*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

(72) Inventors; and

(75) Inventors/Applicants (*for US only*): **KAZEMI, Amir** [US/US]; 53 Southwind Drive, Wallingford, CT US (US). **KOSHY, Beena** [IN/US]; Apartment 11B, 1298 Hartford Turnpike, North Haven, CT 06473 (US). **SANCHIS, Angela** [US/US]; Apartment 4, 7 Trumbull Street, New Haven, CT 06511 (US).

(54) Title: HAPLOTYPES OF THE EDG4 GENE

(57) Abstract: Novel genetic variants of the Endothelial Differentiation, Lysophosphatidic Acid G-Protein-Coupled Receptor, 4 (EDG4) gene are described. Various genotypes, haplotypes, and haplotype pairs that exist in the general United States population are disclosed for the EDG4 gene. Compositions and methods for haplotyping and/or genotyping the EDG4 gene in an individual are also disclosed. Polynucleotides defined by the haplotypes disclosed herein are also described.



**WO 02/12342 A2**

## HAPLOTYPES OF THE EDG4 GENE

## RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application Serial No. 60/223,177 filed August 4, 2000.

## FIELD OF THE INVENTION

This invention relates to variation in genes that encode pharmaceutically-important proteins. In particular, this invention provides genetic variants of the human endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) gene and methods for identifying which variant(s) of this gene is/are possessed by an individual.

## BACKGROUND OF THE INVENTION

Current methods for identifying pharmaceuticals to treat disease often start by identifying, cloning, and expressing an important target protein related to the disease. A determination of whether an agonist or antagonist is needed to produce an effect that may benefit a patient with the disease is then made. Then, vast numbers of compounds are screened against the target protein to find new potential drugs. The desired outcome of this process is a lead compound that is specific for the target, thereby reducing the incidence of the undesired side effects usually caused by activity at non-intended targets. The lead compound identified in this screening process then undergoes further *in vitro* and *in vivo* testing to determine its absorption, disposition, metabolism and toxicological profiles. Typically, this testing involves use of cell lines and animal models with limited, if any, genetic diversity.

What this approach fails to consider, however, is that natural genetic variability exists between individuals in any and every population with respect to pharmaceutically-important proteins, including the protein targets of candidate drugs, the enzymes that metabolize these drugs and the proteins whose activity is modulated by such drug targets. Subtle alteration(s) in the primary nucleotide sequence of a gene encoding a pharmaceutically-important protein may be manifested as significant variation in expression, structure and/or function of the protein. Such alterations may explain the relatively high degree of uncertainty inherent in the treatment of individuals with a drug whose design is based upon a single representative example of the target or enzyme(s) involved in metabolizing the drug. For example, it is well-established that some drugs frequently have lower efficacy in some individuals than others, which means such individuals and their physicians must weigh the possible benefit of a larger dosage against a greater risk of side effects. Also, there is significant variation in how well people metabolize drugs and other exogenous chemicals, resulting in substantial interindividual variation in the toxicity and/or efficacy of such exogenous substances (Evans et al., 1999, *Science* 286:487-491). This variability in efficacy or toxicity of a drug in genetically-diverse patients makes many drugs ineffective or even dangerous in certain groups of the population, leading to the failure of such drugs in clinical

trials or their early withdrawal from the market even though they could be highly beneficial for other groups in the population. This problem significantly increases the time and cost of drug discovery and development, which is a matter of great public concern.

It is well-recognized by pharmaceutical scientists that considering the impact of the genetic variability of pharmaceutically-important proteins in the early phases of drug discovery and development is likely to reduce the failure rate of candidate and approved drugs (Marshall A 1997 *Nature Biotech* 15:1249-52; Kleyn PW et al. 1998 *Science* 281: 1820-21; Kola I 1999 *Curr Opin Biotech* 10:589-92; Hill AVS et al. 1999 in *Evolution in Health and Disease* Stearns SS (Ed.) Oxford University Press, New York, pp 62-76; Meyer U.A. 1999 in *Evolution in Health and Disease* Stearns SS (Ed.) Oxford University Press, New York, pp 41-49; Kalow W et al. 1999 *Clin. Pharm. Therap.* 66:445-7; Marshall, E 1999 *Science* 284:406-7; Judson R et al. 2000 *Pharmacogenomics* 1:1-12; Roses AD 2000 *Nature* 405:857-65). However, in practice this has been difficult to do, in large part because of the time and cost required for discovering the amount of genetic variation that exists in the population (Chakravarti A 1998 *Nature Genet* 19:216-7; Wang DG et al 1998 *Science* 280:1077-82; Chakravarti A 1999 *Nat Genet* 21:56-60 (suppl); Stephens JC 1999 *Mol. Diagnosis* 4:309-317; Kwok PY and Gu S 1999 *Mol. Med. Today* 5:538-43; Davidson S 2000 *Nature Biotech* 18:1134-5).

The standard for measuring genetic variation among individuals is the haplotype, which is the ordered combination of polymorphisms in the sequence of each form of a gene that exists in the population. Because haplotypes represent the variation across each form of a gene, they provide a more accurate and reliable measurement of genetic variation than individual polymorphisms. For example, while specific variations in gene sequences have been associated with a particular phenotype such as disease susceptibility (Roses AD *supra*; Ulbrecht M et al. 2000 *Am J Respir Crit Care Med* 161: 469-74) and drug response (Wolfe CR et al. 2000 *BMJ* 320:987-90; Dahl BS 1997 *Acta Psychiatr Scand* 96 (Suppl 391): 14-21), in many other cases an individual polymorphism may be found in a variety of genomic backgrounds, i.e., different haplotypes, and therefore shows no definitive coupling between the polymorphism and the causative site for the phenotype (Clark AG et al. 1998 *Am J Hum Genet* 63:595-612; Ulbrecht M et al. 2000 *supra*; Drysdale et al. 2000 *PNAS* 97:10483-10488). Thus, there is an unmet need in the pharmaceutical industry for information on what haplotypes exist in the population for pharmaceutically-important genes. Such haplotype information would be useful in improving the efficiency and output of several steps in the drug discovery and development process, including target validation, identifying lead compounds, and early phase clinical trials (Marshall et al., *supra*).

One pharmaceutically-important gene for the treatment of ovarian cancer is the endothelial differentiation, lysophosphatidic acid G protein-coupled receptor, 4 (EDG4) gene or its encoded product. EDG4 is a G protein-coupled receptor involved in mediating the actions of lysophosphatidic acid (LPA), a lysolipid generated by the cleavage of phospholipids from stimulated cells and platelets. LPA mediates signal transduction in cells by binding to a class of G protein-coupled receptors known as EDG receptors, of which EDG4 is a member. LPA is present at micromolar concentrations in serum,

and elicits a variety of biological functions, including proliferation, platelet aggregation, smooth muscle contraction, inhibition of neuroblastoma cell differentiation, chemotaxis and tumor cell invasion (OMIM Entry: 605110). An et al. (*Mol Pharmacol* 1998; 54(5):881-8) determined that LPA increased calcium mobilization in cells expressing EDG4, most likely through inositol triphosphate generated by phospholipase C activation. EDG4 expression has been demonstrated in peripheral blood leukocytes, thymus, spleen, and all tumor types, as well as in leukocytes, testis, prostate, pancreas, and adherent cancer cell lines. High levels of LPA in plasma and ascitic fluid of patients with ovarian cancer suggest the importance of EDG receptors in ovarian cancer cells (OCCs). Specifically, EDG4 was found to evoke proliferative responses in studies of OCCs (Goetzel et al. *Cancer Res* 1999;59(20):5370-5). Thus, EDG4 may represent a distinctive marker of OCC that transduces growth-promoting signals from the high local concentrations of LPA characteristic of aggressive ovarian cancer.

The endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 gene is located on chromosome 19p12 and contains 2 exons that encode a 351 amino acid protein. A reference sequence for the EDG4 gene is shown in the contiguous lines of Figure 1 (GenBank Reference No. 1721380; SEQ ID NO: 1). Reference sequences for the coding sequence (GenBank Accession No. NM\_004720.2) and protein are shown in Figures 2 (SEQ ID NO: 2) and 3 (SEQ ID NO: 3), respectively.

Because of the potential for variation in the EDG4 gene to affect the expression and function of the encoded protein, it would be useful to know whether polymorphisms exist in the EDG4 gene, as well as how such polymorphisms are combined in different copies of the gene. Such information could be applied for studying the biological function of EDG4 as well as in identifying drugs targeting this protein for the treatment of disorders related to its abnormal expression or function.

#### SUMMARY OF THE INVENTION

Accordingly, the inventors herein have discovered 8 novel polymorphic sites in the EDG4 gene. These polymorphic sites (PS) correspond to the following nucleotide positions in Figure 1: 216 (PS1), 395 (PS2), 523 (PS3), 806 (PS4), 812 (PS5), 1291 (PS6), 3099 (PS7) and 3449 (PS8). The polymorphisms at these sites are thymine or cytosine at PS1, cytosine or thymine at PS2, thymine or guanine at PS3, guanine or adenine at PS4, guanine or adenine at PS5, guanine or thymine at PS6, guanine or thymine at PS7 and cytosine or thymine at PS8. In addition, the inventors have determined the identity of the alleles at these sites in a human reference population of 79 unrelated individuals self-identified as belonging to one of four major population groups: African descent, Asian, Caucasian and Hispanic/Latino. From this information, the inventors deduced a set of haplotypes and haplotype pairs for PS1-PS8 in the EDG4 gene, which are shown below in Tables 4 and 3, respectively. Each of these EDG4 haplotypes defines a naturally-occurring isoform (also referred to herein as an "isogene") of the EDG4 gene that exists in the human population. The frequency with which each haplotype and haplotype pair occurs within the total reference population and within each of the four major population

groups included in the reference population was also determined.

Thus, in one embodiment, the invention provides a method, composition and kit for genotyping the EDG4 gene in an individual. The genotyping method comprises identifying the nucleotide pair that is present at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7 and PS8 in both copies of the EDG4 gene from the individual. A genotyping composition of the invention comprises an oligonucleotide probe or primer which is designed to specifically hybridize to a target region containing, or adjacent to, one of these novel EDG4 polymorphic sites. A genotyping kit of the invention comprises a set of oligonucleotides designed to genotype each of these novel EDG4 polymorphic sites. The genotyping method, composition, and kit are useful in determining whether an individual has one of the haplotypes in Table 4 below or has one of the haplotype pairs in Table 3 below.

The invention also provides a method for haplotyping the EDG4 gene in an individual. In one embodiment, the haplotyping method comprises determining, for one copy of the EDG4 gene, the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7 and PS8. In another embodiment, the haplotyping method comprises determining whether one copy of the individual's EDG4 gene is defined by one of the EDG4 haplotypes shown in Table 4, below, or a sub-haplotype thereof. In a preferred embodiment, the haplotyping method comprises determining whether both copies of the individual's EDG4 gene are defined by one of the EDG4 haplotype pairs shown in Table 3 below, or a sub-haplotype pair thereof. The method for establishing the EDG4 haplotype or haplotype pair of an individual is useful for improving the efficiency and reliability of several steps in the discovery and development of drugs for treating diseases associated with EDG4 activity, e.g., ovarian cancer.

For example, the haplotyping method can be used by the pharmaceutical research scientist to validate EDG4 as a candidate target for treating a specific condition or disease predicted to be associated with EDG4 activity. Determining for a particular population the frequency of one or more of the individual EDG4 haplotypes or haplotype pairs described herein will facilitate a decision on whether to pursue EDG4 as a target for treating the specific disease of interest. In particular, if variable EDG4 activity is associated with the disease, then one or more EDG4 haplotypes or haplotype pairs will be found at a higher frequency in disease cohorts than in appropriately genetically matched controls. Conversely, if each of the observed EDG4 haplotypes are of similar frequencies in the disease and control groups, then it may be inferred that variable EDG4 activity has little, if any, involvement with that disease. In either case, the pharmaceutical research scientist can, without *a priori* knowledge as to the phenotypic effect of any EDG4 haplotype or haplotype pair, apply the information derived from detecting EDG4 haplotypes in an individual to decide whether modulating EDG4 activity would be useful in treating the disease.

The claimed invention is also useful in screening for compounds targeting EDG4 to treat a specific condition or disease predicted to be associated with EDG4 activity. For example, detecting

which of the EDG4 haplotypes or haplotype pairs disclosed herein are present in individual members of a population with the specific disease of interest enables the pharmaceutical scientist to screen for a compound(s) that displays the highest desired agonist or antagonist activity for each of the most frequent EDG4 isoforms present in the disease population. Thus, without requiring any *a priori* knowledge of the phenotypic effect of any particular EDG4 haplotype or haplotype pair, the claimed haplotyping method provides the scientist with a tool to identify lead compounds that are more likely to show efficacy in clinical trials.

The method for haplotyping the EDG4 gene in an individual is also useful in the design of clinical trials of candidate drugs for treating a specific condition or disease predicted to be associated with EDG4 activity. For example, instead of randomly assigning patients with the disease of interest to the treatment or control group as is typically done now, determining which of the EDG4 haplotype(s) disclosed herein are present in individual patients enables the pharmaceutical scientist to distribute EDG4 haplotypes and/or haplotype pairs evenly to treatment and control groups, thereby reducing the potential for bias in the results that could be introduced by a larger frequency of an EDG4 haplotype or haplotype pair that had a previously unknown association with response to the drug being studied in the trial. Thus, by practicing the claimed invention, the scientist can more confidently rely on the information learned from the trial, without first determining the phenotypic effect of any EDG4 haplotype or haplotype pair.

In another embodiment, the invention provides a method for identifying an association between a trait and an EDG4 genotype, haplotype, or haplotype pair for one or more of the novel polymorphic sites described herein. The method comprises comparing the frequency of the EDG4 genotype, haplotype, or haplotype pair in a population exhibiting the trait with the frequency of the EDG4 genotype or haplotype in a reference population. A higher frequency of the EDG4 genotype, haplotype, or haplotype pair in the trait population than in the reference population indicates the trait is associated with the EDG4 genotype, haplotype, or haplotype pair. In preferred embodiments, the trait is susceptibility to a disease, severity of a disease, the staging of a disease or response to a drug. In a particularly preferred embodiment, the EDG4 haplotype is selected from the haplotypes shown in Table 4, or a sub-haplotype thereof. Such methods have applicability in developing diagnostic tests and therapeutic treatments for ovarian cancer.

In yet another embodiment, the invention provides an isolated polynucleotide comprising a nucleotide sequence which is a polymorphic variant of a reference sequence for the EDG4 gene or a fragment thereof. The reference sequence comprises the contiguous sequences shown in Figure 1 and the polymorphic variant comprises at least one polymorphism selected from the group consisting of cytosine at PS1, thymine at PS2, guanine at PS3, adenine at PS4, adenine at PS5, thymine at PS6, thymine at PS7 and thymine at PS8.

A particularly preferred polymorphic variant is an isogene of the EDG4 gene. An EDG4 isogene of the invention comprises thymine or cytosine at PS1, cytosine or thymine at PS2, thymine or

guanine at PS3, guanine or adenine at PS4, guanine or adenine at PS5, guanine or thymine at PS6, guanine or thymine at PS7 and cytosine or thymine at PS8. The invention also provides a collection of EDG4 isogenes, referred to herein as an EDG4 genome anthology.

In another embodiment, the invention provides a polynucleotide comprising a polymorphic variant of a reference sequence for an EDG4 cDNA or a fragment thereof. The reference sequence comprises SEQ ID NO:2 (Fig.2) and the polymorphic cDNA comprises at least one polymorphism selected from the group consisting of guanine at a position corresponding to nucleotide 125, adenine at a position corresponding to nucleotide 408 and adenine at a position corresponding to nucleotide 414. A particularly preferred polymorphic cDNA variant comprises the coding sequence of an EDG4 isogene defined by haplotypes 1c-4c and 7c.

Polynucleotides complementary to these EDG4 genomic and cDNA variants are also provided by the invention. It is believed that polymorphic variants of the EDG4 gene will be useful in studying the expression and function of EDG4, and in expressing EDG4 protein for use in screening for candidate drugs to treat diseases related to EDG4 activity.

In other embodiments, the invention provides a recombinant expression vector comprising one of the polymorphic genomic variants operably linked to expression regulatory elements as well as a recombinant host cell transformed or transfected with the expression vector. The recombinant vector and host cell may be used to express EDG4 for protein structure analysis and drug binding studies.

In yet another embodiment, the invention provides a polypeptide comprising a polymorphic variant of a reference amino acid sequence for the EDG4 protein. The reference amino acid sequence comprises SEQ ID NO:3 (Fig.3) and the polymorphic variant comprises glycine at a position corresponding to amino acid position 42. A polymorphic variant of EDG4 is useful in studying the effect of the variation on the biological activity of EDG4 as well as on the binding affinity of candidate drugs targeting EDG4 for the treatment of ovarian cancer.

The present invention also provides antibodies that recognize and bind to the above polymorphic EDG4 protein variant. Such antibodies can be utilized in a variety of diagnostic and prognostic formats and therapeutic methods.

The present invention also provides nonhuman transgenic animals comprising one of the EDG4 polymorphic genomic variants described herein and methods for producing such animals. The transgenic animals are useful for studying expression of the EDG4 isogenes *in vivo*, for *in vivo* screening and testing of drugs targeted against EDG4 protein, and for testing the efficacy of therapeutic agents and compounds for ovarian cancer in a biological system.

The present invention also provides a computer system for storing and displaying polymorphism data determined for the EDG4 gene. The computer system comprises a computer processing unit; a display; and a database containing the polymorphism data. The polymorphism data includes the polymorphisms, the genotypes and the haplotypes identified for the EDG4 gene in a reference population. In a preferred embodiment, the computer system is capable of producing a

display showing EDG4 haplotypes organized according to their evolutionary relationships.

#### BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a reference sequence for the EDG4 gene (Genaissance Reference No.

1721380; contiguous lines), with the start and stop positions of each region of coding sequence indicated with a bracket ([ or ]) and the numerical position below the sequence and the polymorphic site(s) and polymorphism(s) identified by Applicants in a reference population indicated by the variant nucleotide positioned below the polymorphic site in the sequence. SEQ ID NO:1 is equivalent to

Figure 1, with the two alternative allelic variants of each polymorphic site indicated by the appropriate

nucleotide symbol (R= G or A, Y= T or C, M= A or C, K= G or T, S= G or C, and W= A or T; WIPO standard ST.25). SEQ ID NO:44 is a modified version of SEQ ID NO:1 that shows the context sequence of each polymorphic site, PS1-PS8, in a uniform format to facilitate electronic searching. For

each polymorphic site, SEQ ID NO:44 contains a block of 60 bases of the nucleotide sequence

encompassing the centrally-located polymorphic site at the 30<sup>th</sup> position, followed by 60 bases of

unspecified sequence to represent that each PS is separated by genomic sequence whose composition is defined elsewhere herein.

Figure 2 illustrates a reference sequence for the EDG4 coding sequence (contiguous lines; SEQ ID NO:2), with the polymorphic site(s) and polymorphism(s) identified by Applicants in a reference population indicated by the variant nucleotide positioned below the polymorphic site in the sequence.

Figure 3 illustrates a reference sequence for the EDG4 protein (contiguous lines; SEQ ID NO:3), with the variant amino acid(s) caused by the polymorphism(s) of Figure 2 positioned below the polymorphic site in the sequence.

#### DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention is based on the discovery of novel variants of the EDG4 gene. As described in more detail below, the inventors herein discovered 8 isogenes of the EDG4 gene by characterizing the EDG4 gene found in genomic DNAs isolated from an Index Repository that contains immortalized cell lines from one chimpanzee and 93 human individuals. The human individuals included a reference population of 79 unrelated individuals self-identified as belonging to one of four major population groups: Caucasian (21 individuals), African descent (20 individuals), Asian (20 individuals), or Hispanic/Latino (18 individuals). To the extent possible, the members of this reference population were organized into population subgroups by their self-identified ethnogeographic origin as shown in Table 1 below.



Table 1. Population Groups in the Index Repository

Population Group	Population Subgroup	No. of Individuals
African descent		20
	Sierra Leone	1
Asian		20
	Burma	1
	China	3
	Japan	6
	Korea	1
	Philippines	5
	Vietnam	4
Caucasian		21
	British Isles	3
	British Isles/Central	4
	British Isles/Eastern	1
	Central/Eastern	1
	Eastern	3
	Central/Mediterranean	1
	Mediterranean	2
	Scandinavian	2
Hispanic/Latino		18
	Caribbean	8
	Caribbean (Spanish Descent)	2
	Central American (Spanish Descent)	1
	Mexican American	4
	South American (Spanish Descent)	3

In addition, the Index Repository contains three unrelated indigenous American Indians (one from each of North, Central and South America), one three-generation Caucasian family (from the CEPH Utah cohort) and one two-generation African-American family.

The EDG4 isogenes present in the human reference population are defined by haplotypes for 8 polymorphic sites in the EDG4 gene, all of which are believed to be novel. The novel EDG4 polymorphic sites identified by the inventors are referred to as PS1-PS8 to designate the order in which they are located in the gene (see Table 2 below). Using the genotypes identified in the Index Repository for PS1-PS8 and the methodology described in the Examples below, the inventors herein also determined the pair of haplotypes for the EDG4 gene present in individual human members of this repository. The human genotypes and haplotypes found in the repository for the EDG4 gene include those shown in Tables 3 and 4, respectively. The polymorphism and haplotype data disclosed herein are useful for validating whether EDG4 is a suitable target for drugs to treat ovarian cancer, screening for such drugs and reducing bias in clinical trials of such drugs.

In the context of this disclosure, the following terms shall be defined as follows unless otherwise indicated:

**Allele** - A particular form of a genetic locus, distinguished from other forms by its particular nucleotide sequence.

**Candidate Gene** - A gene which is hypothesized to be responsible for a disease, condition, or

the response to a treatment, or to be correlated with one of these.

**Gene** - A segment of DNA that contains all the information for the regulated biosynthesis of an RNA product, including promoters, exons, introns, and other untranslated regions that control expression.

5       **Genotype** – An unphased 5' to 3' sequence of nucleotide pair(s) found at one or more polymorphic sites in a locus on a pair of homologous chromosomes in an individual. As used herein, genotype includes a full-genotype and/or a sub-genotype as described below.

**Full-genotype** – The unphased 5' to 3' sequence of nucleotide pairs found at all polymorphic sites examined herein in a locus on a pair of homologous chromosomes in a single individual.

10       **Sub-genotype** – The unphased 5' to 3' sequence of nucleotides seen at a subset of the polymorphic sites examined herein in a locus on a pair of homologous chromosomes in a single individual.

**Genotyping** – A process for determining a genotype of an individual.

**Haplotype** – A 5' to 3' sequence of nucleotides found at one or more polymorphic sites in a  
15       locus on a single chromosome from a single individual. As used herein, haplotype includes a full-haplotype and/or a sub-haplotype as described below.

**Full-haplotype** – The 5' to 3' sequence of nucleotides found at all polymorphic sites examined herein in a locus on a single chromosome from a single individual.

**Sub-haplotype** – The 5' to 3' sequence of nucleotides seen at a subset of the polymorphic sites  
20       examined herein in a locus on a single chromosome from a single individual.

**Haplotype pair** – The two haplotypes found for a locus in a single individual.

**Haplotyping** – A process for determining one or more haplotypes in an individual and includes use of family pedigrees, molecular techniques and/or statistical inference.

**Haplotype data** - Information concerning one or more of the following for a specific gene: a  
25       listing of the haplotype pairs in each individual in a population; a listing of the different haplotypes in a population; frequency of each haplotype in that or other populations, and any known associations between one or more haplotypes and a trait.

**Isoform** – A particular form of a gene, mRNA, cDNA or the protein encoded thereby, distinguished from other forms by its particular sequence and/or structure.

30       **Isogene** – One of the isoforms of a gene found in a population. An isogene contains all of the polymorphisms present in the particular isoform of the gene.

**Isolated** – As applied to a biological molecule such as RNA, DNA, oligonucleotide, or protein, isolated means the molecule is substantially free of other biological molecules such as nucleic acids, proteins, lipids, carbohydrates, or other material such as cellular debris and growth media. Generally,  
35       the term "isolated" is not intended to refer to a complete absence of such material or to absence of water, buffers, or salts, unless they are present in amounts that substantially interfere with the methods of the present invention.

**Locus** – A location on a chromosome or DNA molecule corresponding to a gene or a physical or phenotypic feature, wherein the physical feature may be a polymorphic site.

**Naturally-occurring** – A term used to designate that the object it is applied to, e.g., naturally-occurring polynucleotide or polypeptide, can be isolated from a source in nature and which has not been intentionally modified by man.

**Nucleotide pair** – The nucleotides found at a polymorphic site on the two copies of a chromosome from an individual.

**Phased** – As applied to a sequence of nucleotide pairs for two or more polymorphic sites in a locus, phased means the combination of nucleotides present at those polymorphic sites on a single copy of the locus is known.

**Polymorphic site (PS)** – A position within a locus at which at least two alternative sequences are found in a population.

**Polymorphic variant** – A gene, mRNA, cDNA, polypeptide or peptide whose nucleotide or amino acid sequence varies from a reference sequence due to the presence of a polymorphism in the gene.

**Polymorphism** – The sequence variation observed in an individual at a polymorphic site. Polymorphisms include nucleotide substitutions, insertions, deletions and microsatellites and may, but need not, result in detectable differences in gene expression or protein function.

**Polymorphism data** – Information concerning one or more of the following for a specific gene: location of polymorphic sites; sequence variation at those sites; frequency of polymorphisms in one or more populations; the different genotypes and/or haplotypes determined for the gene; frequency of one or more of these genotypes and/or haplotypes in one or more populations; any known association(s) between a trait and a genotype or a haplotype for the gene.

**Polymorphism Database** – A collection of polymorphism data arranged in a systematic or methodical way and capable of being individually accessed by electronic or other means.

**Polynucleotide** – A nucleic acid molecule comprised of single-stranded RNA or DNA or comprised of complementary, double-stranded DNA.

**Population Group** – A group of individuals sharing a common ethnogeographic origin.

**Reference Population** – A group of subjects or individuals who are predicted to be representative of the genetic variation found in the general population. Typically, the reference population represents the genetic variation in the population at a certainty level of at least 85%, preferably at least 90%, more preferably at least 95% and even more preferably at least 99%.

**Single Nucleotide Polymorphism (SNP)** – Typically, the specific pair of nucleotides observed at a single polymorphic site. In rare cases, three or four nucleotides may be found.

**Subject** – A human individual whose genotypes or haplotypes or response to treatment or disease state are to be determined.

**Treatment** – A stimulus administered internally or externally to a subject.

**Unphased** – As applied to a sequence of nucleotide pairs for two or more polymorphic sites in a locus, unphased means the combination of nucleotides present at those polymorphic sites on a single copy of the locus is not known.

As discussed above, information on the identity of genotypes and haplotypes for the EDG4 gene of any particular individual as well as the frequency of such genotypes and haplotypes in any particular population of individuals is expected to be useful for a variety of drug discovery and development applications. Thus, the invention also provides compositions and methods for detecting the novel EDG4 polymorphisms and haplotypes identified herein.

The compositions comprise at least one EDG4 genotyping oligonucleotide. In one embodiment, an EDG4 genotyping oligonucleotide is a probe or primer capable of hybridizing to a target region that is located close to, or that contains, one of the novel polymorphic sites described herein. As used herein, the term “oligonucleotide” refers to a polynucleotide molecule having less than about 100 nucleotides. A preferred oligonucleotide of the invention is 10 to 35 nucleotides long. More preferably, the oligonucleotide is between 15 and 30, and most preferably, between 20 and 25 nucleotides in length. The exact length of the oligonucleotide will depend on many factors that are routinely considered and practiced by the skilled artisan. The oligonucleotide may be comprised of any phosphorylation state of ribonucleotides, deoxyribonucleotides, and acyclic nucleotide derivatives, and other functionally equivalent derivatives. Alternatively, oligonucleotides may have a phosphate-free backbone, which may be comprised of linkages such as carboxymethyl, acetamidate, carbamate, polyamide (peptide nucleic acid (PNA)) and the like (Varma, R. in Molecular Biology and Biotechnology, A Comprehensive Desk Reference, Ed. R. Meyers, VCH Publishers, Inc. (1995), pages 617-620). Oligonucleotides of the invention may be prepared by chemical synthesis using any suitable methodology known in the art, or may be derived from a biological sample, for example, by restriction digestion. The oligonucleotides may be labeled, according to any technique known in the art, including use of radiolabels, fluorescent labels, enzymatic labels, proteins, haptens, antibodies, sequence tags and the like.

Genotyping oligonucleotides of the invention must be capable of specifically hybridizing to a target region of an EDG4 polynucleotide, i.e., an EDG4 isogene. As used herein, specific hybridization means the oligonucleotide forms an anti-parallel double-stranded structure with the target region under certain hybridizing conditions, while failing to form such a structure when incubated with a non-target region or a non-EDG4 polynucleotide under the same hybridizing conditions. Preferably, the oligonucleotide specifically hybridizes to the target region under conventional high stringency conditions. The skilled artisan can readily design and test oligonucleotide probes and primers suitable for detecting polymorphisms in the EDG4 gene using the polymorphism information provided herein in conjunction with the known sequence information for the EDG4 gene and routine techniques.

A nucleic acid molecule such as an oligonucleotide or polynucleotide is said to be a “perfect” or “complete” complement of another nucleic acid molecule if every nucleotide of one of the molecules is

complementary to the nucleotide at the corresponding position of the other molecule. A nucleic acid molecule is "substantially complementary" to another molecule if it hybridizes to that molecule with sufficient stability to remain in a duplex form under conventional low-stringency conditions.

Conventional hybridization conditions are described, for example, by Sambrook J. et al., in *Molecular Cloning, A Laboratory Manual*, 2<sup>nd</sup> Edition, Cold Spring Harbor Press, Cold Spring Harbor, NY (1989) and by Haymes, B.D. et al. in *Nucleic Acid Hybridization, A Practical Approach*, IRL Press, Washington, D.C. (1985). While perfectly complementary oligonucleotides are preferred for detecting polymorphisms, departures from complete complementarity are contemplated where such departures do not prevent the molecule from specifically hybridizing to the target region. For example, an oligonucleotide primer may have a non-complementary fragment at its 5' end, with the remainder of the primer being complementary to the target region. Alternatively, non-complementary nucleotides may be interspersed into the oligonucleotide probe or primer as long as the resulting probe or primer is still capable of specifically hybridizing to the target region.

Preferred genotyping oligonucleotides of the invention are allele-specific oligonucleotides. As used herein, the term allele-specific oligonucleotide (ASO) means an oligonucleotide that is able, under sufficiently stringent conditions, to hybridize specifically to one allele of a gene, or other locus, at a target region containing a polymorphic site while not hybridizing to the corresponding region in another allele(s). As understood by the skilled artisan, allele-specificity will depend upon a variety of readily optimized stringency conditions, including salt and formamide concentrations, as well as temperatures for both the hybridization and washing steps. Examples of hybridization and washing conditions typically used for ASO probes are found in Kogan et al., "Genetic Prediction of Hemophilia A" in *PCR Protocols, A Guide to Methods and Applications*, Academic Press, 1990 and Ruaño et al., 87 *Proc. Natl. Acad. Sci. USA* 6296-6300, 1990. Typically, an ASO will be perfectly complementary to one allele while containing a single mismatch for another allele.

Allele-specific oligonucleotides of the invention include ASO probes and ASO primers. ASO probes which usually provide good discrimination between different alleles are those in which a central position of the oligonucleotide probe aligns with the polymorphic site in the target region (e.g., approximately the 7<sup>th</sup> or 8<sup>th</sup> position in a 15mer, the 8<sup>th</sup> or 9<sup>th</sup> position in a 16mer, and the 10<sup>th</sup> or 11<sup>th</sup> position in a 20mer). An ASO primer of the invention has a 3' terminal nucleotide, or preferably a 3' penultimate nucleotide, that is complementary to only one nucleotide of a particular SNP, thereby acting as a primer for polymerase-mediated extension only if the allele containing that nucleotide is present. ASO probes and primers hybridizing to either the coding or noncoding strand are contemplated by the invention.

ASO probes and primers listed below use the appropriate nucleotide symbol (R= G or A, Y= T or C, M= A or C, K= G or T, S= G or C, and W= A or T; WIPO standard ST.25) at the position of the polymorphic site to represent that the ASO contains one of the two alternative allelic variants observed at that polymorphic site.

A preferred ASO probe for detecting EDG4 gene polymorphisms comprises a nucleotide sequence, listed 5' to 3', selected from the group consisting of:

- 5 CTCCACCYCCACCCC (SEQ ID NO:4) and its complement,  
 TGTGTCCYAGATGG (SEQ ID NO:5) and its complement,  
 CTGACCGKCAGCGTG (SEQ ID NO:6) and its complement,  
 TGGCCGTRCAGCTGC (SEQ ID NO:7) and its complement,  
 TGCAGCTRCACAGCC (SEQ ID NO:8) and its complement,  
 AATGTATKTGAGCAG (SEQ ID NO:9) and its complement,  
 10 CACTTCCCKCTCTGT (SEQ ID NO:10) and its complement, and  
 AGCGGTAYGCGGCAA (SEQ ID NO:11), and its complement.

A preferred ASO primer for detecting EDG4 gene polymorphisms comprises a nucleotide sequence, listed 5' to 3', selected from the group consisting of:

- 15 CTGCTCCTCCACCYC (SEQ ID NO:12); GGGGCTGGGGTGGRG (SEQ ID NO:13);  
 ATCTCTTGTGTCCYC (SEQ ID NO:14); TGATGACCATCTGRG (SEQ ID NO:15);  
 CTGGGGCTGACCGKC (SEQ ID NO:16); CACCAGCAGCTGMC (SEQ ID NO:17);  
 GTGTGATGGCCGTRC (SEQ ID NO:18); GGCTGTGCAGCTGYA (SEQ ID NO:19);  
 20 TGGCCGTGCAGCTRC (SEQ ID NO:20); GCAGGCGGCTGTGYA (SEQ ID NO:21);  
 GGTGATAATGTATKT (SEQ ID NO:22); AGGCCTCTGCTCAMA (SEQ ID NO:23);  
 CCTGCTCACTTCCKC (SEQ ID NO:24); TGTGGACAGAGGMG (SEQ ID NO:25);  
 AACTTCAGCGGTAYG (SEQ ID NO:26); and TGTGCTTGCCGCRT (SEQ ID NO:27).

- 25 Other genotyping oligonucleotides of the invention hybridize to a target region located one to several nucleotides downstream of one of the novel polymorphic sites identified herein. Such oligonucleotides are useful in polymerase-mediated primer extension methods for detecting one of the novel polymorphisms described herein and therefore such genotyping oligonucleotides are referred to herein as "primer-extension oligonucleotides". In a preferred embodiment, the 3'-terminus of a primer-  
 30 extension oligonucleotide is a deoxynucleotide complementary to the nucleotide located immediately adjacent to the polymorphic site.

A particularly preferred oligonucleotide primer for detecting EDG4 gene polymorphisms by primer extension terminates in a nucleotide sequence, listed 5' to 3', selected from the group consisting of:

- 35 CTCCTCCACC (SEQ ID NO:28); GCTGGGGTGG (SEQ ID NO:29);  
 TCTTGTGTCC (SEQ ID NO:30); TGACCATCTG (SEQ ID NO:31);  
 GGGCTGACCG (SEQ ID NO:32); CAGCACGCTG (SEQ ID NO:33);  
 TGATGGCCGT (SEQ ID NO:34); TGTGCAGCTG (SEQ ID NO:35);  
 40 CCGTGCAGCT (SEQ ID NO:36); GCGGCTGTG (SEQ ID NO:37);  
 GATAATGTAT (SEQ ID NO:38); CCTCTGCTCA (SEQ ID NO:39);  
 GCTCACTTCC (SEQ ID NO:40); GGGACAGAGG (SEQ ID NO:41);  
 TTCAGCGGTA (SEQ ID NO:42); and TGCTTGCCGC (SEQ ID NO:43).

- 45 In some embodiments, a composition contains two or more differently labeled genotyping oligonucleotides for simultaneously probing the identity of nucleotides at two or more polymorphic sites. It is also contemplated that primer compositions may contain two or more sets of allele-specific

primer pairs to allow simultaneous targeting and amplification of two or more regions containing a polymorphic site.

EDG4 genotyping oligonucleotides of the invention may also be immobilized on or synthesized on a solid surface such as a microchip, bead, or glass slide (see, e.g., WO 98/20020 and WO 98/20019).

5 Such immobilized genotyping oligonucleotides may be used in a variety of polymorphism detection assays, including but not limited to probe hybridization and polymerase extension assays. Immobilized EDG4 genotyping oligonucleotides of the invention may comprise an ordered array of oligonucleotides designed to rapidly screen a DNA sample for polymorphisms in multiple genes at the same time.

10 In another embodiment, the invention provides a kit comprising at least two genotyping oligonucleotides packaged in separate containers. The kit may also contain other components such as hybridization buffer (where the oligonucleotides are to be used as a probe) packaged in a separate container. Alternatively, where the oligonucleotides are to be used to amplify a target region, the kit may contain, packaged in separate containers, a polymerase and a reaction buffer optimized for primer extension mediated by the polymerase, such as PCR.

15 The above described oligonucleotide compositions and kits are useful in methods for genotyping and/or haplotyping the EDG4 gene in an individual. As used herein, the terms "EDG4 genotype" and "EDG4 haplotype" mean the genotype or haplotype contains the nucleotide pair or nucleotide, respectively, that is present at one or more of the novel polymorphic sites described herein and may optionally also include the nucleotide pair or nucleotide present at one or more additional  
20 polymorphic sites in the EDG4 gene. The additional polymorphic sites may be currently known polymorphic sites or sites that are subsequently discovered.

One embodiment of the genotyping method involves isolating from the individual a nucleic acid sample comprising the two copies of the EDG4 gene, or a fragment thereof, that are present in the individual, and determining the identity of the nucleotide pair at one or more polymorphic sites selected  
25 from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7 and PS8 in the two copies to assign an EDG4 genotype to the individual. As will be readily understood by the skilled artisan, the two "copies" of a gene in an individual may be the same allele or may be different alleles. In a particularly preferred embodiment, the genotyping method comprises determining the identity of the nucleotide pair at each of PS1-PS8.

30 Typically, the nucleic acid sample is isolated from a biological sample taken from the individual, such as a blood sample or tissue sample. Suitable tissue samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair. The nucleic acid sample may be comprised of genomic DNA, mRNA, or cDNA and, in the latter two cases, the biological sample must be obtained from a tissue in which the EDG4 gene is expressed. Furthermore it will be understood by  
35 the skilled artisan that mRNA or cDNA preparations would not be used to detect polymorphisms located in introns or in 5' and 3' untranslated regions. If an EDG4 gene fragment is isolated, it must contain the polymorphic site(s) to be genotyped.

One embodiment of the haplotyping method comprises isolating from the individual a nucleic acid sample containing only one of the two copies of the EDG4 gene, or a fragment thereof, that is present in the individual and determining in that copy the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7 and PS8 in that copy to assign an EDG4 haplotype to the individual. The nucleic acid may be isolated using any method capable of separating the two copies of the EDG4 gene or fragment such as one of the methods described above for preparing EDG4 isogenes, with targeted *in vivo* cloning being the preferred approach. As will be readily appreciated by those skilled in the art, any individual clone will only provide haplotype information on one of the two EDG4 gene copies present in an individual. If haplotype information is desired for the individual's other copy, additional EDG4 clones will need to be examined. Typically, at least five clones should be examined to have more than a 90% probability of haplotyping both copies of the EDG4 gene in an individual. In a particularly preferred embodiment, the nucleotide at each of PS1-PS8 is identified.

In another embodiment, the haplotyping method comprises determining whether an individual has one or more of the EDG4 haplotypes shown in Table 4. This can be accomplished by identifying, for one or both copies of the individual's EDG4 gene, the phased sequence of nucleotides present at each of PS1-PS8. The present invention also contemplates that typically only a subset of PS1-PS8 will need to be directly examined to assign to an individual one or more of the haplotypes shown in Table 4. This is because at least one polymorphic site in a gene is frequently in strong linkage disequilibrium with one or more other polymorphic sites in that gene (Drysdales, CM et al. 2000 *PNAS* 97:10483-10488; Rieder MJ et al. 1999 *Nature Genetics* 22:59-62). Two sites are said to be in linkage disequilibrium if the presence of a particular variant at one site enhances the predictability of another variant at the second site (Stephens, JC 1999, *Mol. Diag.* 4:309-317). Techniques for determining whether any two polymorphic sites are in linkage disequilibrium are well-known in the art (Weir B.S. 1996 *Genetic Data Analysis II*, Sinauer Associates, Inc. Publishers, Sunderland, MA).

In a preferred embodiment, an EDG4 haplotype pair is determined for an individual by identifying the phased sequence of nucleotides at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7 and PS8 in each copy of the EDG4 gene that is present in the individual. In a particularly preferred embodiment, the haplotyping method comprises identifying the phased sequence of nucleotides at each of PS1-PS8 in each copy of the EDG4 gene. When haplotyping both copies of the gene, the identifying step is preferably performed with each copy of the gene being placed in separate containers. However, it is also envisioned that if the two copies are labeled with different tags, or are otherwise separately distinguishable or identifiable, it could be possible in some cases to perform the method in the same container. For example, if first and second copies of the gene are labeled with different first and second fluorescent dyes, respectively, and an allele-specific oligonucleotide labeled with yet a third different fluorescent dye is used to assay the polymorphic site(s), then detecting a combination of the first and third dyes would identify the



polymorphism in the first gene copy while detecting a combination of the second and third dyes would identify the polymorphism in the second gene copy.

In both the genotyping and haplotyping methods, the identity of a nucleotide (or nucleotide pair) at a polymorphic site(s) may be determined by amplifying a target region(s) containing the polymorphic site(s) directly from one or both copies of the EDG4 gene, or a fragment thereof, and the sequence of the amplified region(s) determined by conventional methods. It will be readily appreciated by the skilled artisan that only one nucleotide will be detected at a polymorphic site in individuals who are homozygous at that site, while two different nucleotides will be detected if the individual is heterozygous for that site. The polymorphism may be identified directly, known as positive-type identification, or by inference, referred to as negative-type identification. For example, where a SNP is known to be guanine and cytosine in a reference population, a site may be positively determined to be either guanine or cytosine for an individual homozygous at that site, or both guanine and cytosine, if the individual is heterozygous at that site. Alternatively, the site may be negatively determined to be not guanine (and thus cytosine/cytosine) or not cytosine (and thus guanine/guanine).

The target region(s) may be amplified using any oligonucleotide-directed amplification method, including but not limited to polymerase chain reaction (PCR) (U.S. Patent No. 4,965,188), ligase chain reaction (LCR) (Barany et al., *Proc. Natl. Acad. Sci. USA* 88:189-193, 1991; WO90/01069), and oligonucleotide ligation assay (OLA) (Landegren et al., *Science* 241:1077-1080, 1988).

Other known nucleic acid amplification procedures may be used to amplify the target region including transcription-based amplification systems (U.S. Patent No. 5,130,238; EP 329,822; U.S. Patent No. 5,169,766, WO89/06700) and isothermal methods (Walker et al., *Proc. Natl. Acad. Sci. USA* 89:392-396, 1992).

A polymorphism in the target region may also be assayed before or after amplification using one of several hybridization-based methods known in the art. Typically, allele-specific oligonucleotides are utilized in performing such methods. The allele-specific oligonucleotides may be used as differently labeled probe pairs, with one member of the pair showing a perfect match to one variant of a target sequence and the other member showing a perfect match to a different variant. In some embodiments, more than one polymorphic site may be detected at once using a set of allele-specific oligonucleotides or oligonucleotide pairs. Preferably, the members of the set have melting temperatures within 5°C, and more preferably within 2°C, of each other when hybridizing to each of the polymorphic sites being detected.

Hybridization of an allele-specific oligonucleotide to a target polynucleotide may be performed with both entities in solution, or such hybridization may be performed when either the oligonucleotide or the target polynucleotide is covalently or noncovalently affixed to a solid support. Attachment may be mediated, for example, by antibody-antigen interactions, poly-L-Lys, streptavidin or avidin-biotin, salt bridges, hydrophobic interactions, chemical linkages, UV cross-linking baking, etc. Allele-specific oligonucleotides may be synthesized directly on the solid support or attached to the solid support

subsequent to synthesis. Solid-supports suitable for use in detection methods of the invention include substrates made of silicon, glass, plastic, paper and the like, which may be formed, for example, into wells (as in 96-well plates), slides, sheets, membranes, fibers, chips, dishes, and beads. The solid support may be treated, coated or derivatized to facilitate the immobilization of the allele-specific oligonucleotide or target nucleic acid.

The genotype or haplotype for the EDG4 gene of an individual may also be determined by hybridization of a nucleic acid sample containing one or both copies of the gene, or fragment(s) thereof, to nucleic acid arrays and subarrays such as described in WO 95/11995. The arrays would contain a battery of allele-specific oligonucleotides representing each of the polymorphic sites to be included in the genotype or haplotype.

The identity of polymorphisms may also be determined using a mismatch detection technique, including but not limited to the RNase protection method using riboprobes (Winter et al., *Proc. Natl. Acad. Sci. USA* 82:7575, 1985; Meyers et al., *Science* 230:1242, 1985) and proteins which recognize nucleotide mismatches, such as the E. coli mutS protein (Modrich, P. *Ann. Rev. Genet.* 25:229-253, 1991). Alternatively, variant alleles can be identified by single strand conformation polymorphism (SSCP) analysis (Orita et al., *Genomics* 5:874-879, 1989; Humphries et al., in *Molecular Diagnosis of Genetic Diseases*, R. Elles, ed., pp. 321-340, 1996) or denaturing gradient gel electrophoresis (DGGE) (Wartell et al., *Nucl. Acids Res.* 18:2699-2706, 1990; Sheffield et al., *Proc. Natl. Acad. Sci. USA* 86:232-236, 1989).

A polymerase-mediated primer extension method may also be used to identify the polymorphism(s). Several such methods have been described in the patent and scientific literature and include the "Genetic Bit Analysis" method (WO92/15712) and the ligase/polymerase mediated genetic bit analysis (U.S. Patent 5,679,524. Related methods are disclosed in WO91/02087, WO90/09455, WO95/17676, U.S. Patent Nos. 5,302,509, and 5,945,283. Extended primers containing a polymorphism may be detected by mass spectrometry as described in U.S. Patent No. 5,605,798. Another primer extension method is allele-specific PCR (Ruano et al., *Nucl. Acids Res.* 17:8392, 1989; Ruano et al., *Nucl. Acids Res.* 19, 6877-6882, 1991; WO 93/22456; Turki et al., *J. Clin. Invest.* 95:1635-1641, 1995). In addition, multiple polymorphic sites may be investigated by simultaneously amplifying multiple regions of the nucleic acid using sets of allele-specific primers as described in Wallace et al. (WO89/10414).

In addition, the identity of the allele(s) present at any of the novel polymorphic sites described herein may be indirectly determined by genotyping another polymorphic site that is in linkage disequilibrium with the polymorphic site that is of interest. Polymorphic sites in linkage disequilibrium with the presently disclosed polymorphic sites may be located in regions of the gene or in other genomic regions not examined herein. Genotyping of a polymorphic site in linkage disequilibrium with the novel polymorphic sites described herein may be performed by, but is not limited to, any of the above-mentioned methods for detecting the identity of the allele at a polymorphic site.

In another aspect of the invention, an individual's EDG4 haplotype pair is predicted from its EDG4 genotype using information on haplotype pairs known to exist in a reference population. In its broadest embodiment, the haplotyping prediction method comprises identifying an EDG4 genotype for the individual at two or more EDG4 polymorphic sites described herein, enumerating all possible haplotype pairs which are consistent with the genotype, accessing data containing EDG4 haplotype pairs identified in a reference population, and assigning a haplotype pair to the individual that is consistent with the data. In one embodiment, the reference haplotype pairs include the EDG4 haplotype pairs shown in Table 3.

Generally, the reference population should be composed of randomly-selected individuals representing the major ethnogeographic groups of the world. A preferred reference population for use in the methods of the present invention comprises an approximately equal number of individuals from Caucasian, African-descent, Asian and Hispanic-Latino population groups with the minimum number of each group being chosen based on how rare a haplotype one wants to be guaranteed to see. For example, if one wants to have a  $q\%$  chance of not missing a haplotype that exists in the population at a  $p\%$  frequency of occurring in the reference population, the number of individuals ( $n$ ) who must be sampled is given by  $2n = \log(1-q)/\log(1-p)$  where  $p$  and  $q$  are expressed as fractions. A preferred reference population allows the detection of any haplotype whose frequency is at least 10% with about 99% certainty and comprises about 20 unrelated individuals from each of the four population groups named above. A particularly preferred reference population includes a 3-generation family representing one or more of the four population groups to serve as controls for checking quality of haplotyping procedures.

In a preferred embodiment, the haplotype frequency data for each ethnogeographic group is examined to determine whether it is consistent with Hardy-Weinberg equilibrium. Hardy-Weinberg equilibrium (D.L. Hartl et al., *Principles of Population Genomics*, Sinauer Associates (Sunderland, MA), 3<sup>rd</sup> Ed., 1997) postulates that the frequency of finding the haplotype pair  $H_1 / H_2$  is equal to  $p_{H-W}(H_1 / H_2) = 2p(H_1)p(H_2)$  if  $H_1 \neq H_2$  and  $p_{H-W}(H_1 / H_2) = p(H_1)p(H_2)$  if  $H_1 = H_2$ . A statistically significant difference between the observed and expected haplotype frequencies could be due to one or more factors including significant inbreeding in the population group, strong selective pressure on the gene, sampling bias, and/or errors in the genotyping process. If large deviations from Hardy-Weinberg equilibrium are observed in an ethnogeographic group, the number of individuals in that group can be increased to see if the deviation is due to a sampling bias. If a larger sample size does not reduce the difference between observed and expected haplotype pair frequencies, then one may wish to consider haplotyping the individual using a direct haplotyping method such as, for example, CLASPER System<sup>™</sup> technology (U.S. Patent No. 5,866,404), single molecule dilution, or allele-specific long-range PCR (Michalotos-Beloin et al., *Nucleic Acids Res.* 24:4841-4843, 1996).

In one embodiment of this method for predicting an EDG4 haplotype pair for an individual, the

assigning step involves performing the following analysis. First, each of the possible haplotype pairs is compared to the haplotype pairs in the reference population. Generally, only one of the haplotype pairs in the reference population matches a possible haplotype pair and that pair is assigned to the individual. Occasionally, only one haplotype represented in the reference haplotype pairs is consistent with a possible haplotype pair for an individual, and in such cases the individual is assigned a haplotype pair containing this known haplotype and a new haplotype derived by subtracting the known haplotype from the possible haplotype pair. Alternatively, the haplotype pair in an individual may be predicted from the individual's genotype for that gene using reported methods (e.g., Clark et al. 1990 *Mol Bio Evol* 7:111-22) or through a commercial haplotyping service such as offered by Genaissance Pharmaceuticals, Inc. (New Haven, CT). In rare cases, either no haplotypes in the reference population are consistent with the possible haplotype pairs, or alternatively, multiple reference haplotype pairs are consistent with the possible haplotype pairs. In such cases, the individual is preferably haplotyped using a direct molecular haplotyping method such as, for example, CLASPER System™ technology (U.S. Patent No. 5,866,404), SMD, or allele-specific long-range PCR (Michalotos-Beloin et al., *supra*). A preferred process for predicting EDG4 haplotype pairs from EDG4 genotypes is described in U.S. Provisional Application Serial No. 60/198,340 and the corresponding International Application, PCT/US01/12831.

The invention also provides a method for determining the frequency of an EDG4 genotype, haplotype, or haplotype pair in a population. The method comprises, for each member of the population, determining the genotype or the haplotype pair for the novel EDG4 polymorphic sites described herein, and calculating the frequency any particular genotype, haplotype, or haplotype pair is found in the population. The population may be a reference population, a family population, a same sex population, a population group, or a trait population (e.g., a group of individuals exhibiting a trait of interest such as a medical condition or response to a therapeutic treatment).

In another aspect of the invention, frequency data for EDG4 genotypes, haplotypes, and/or haplotype pairs are determined in a reference population and used in a method for identifying an association between a trait and an EDG4 genotype, haplotype, or haplotype pair. The trait may be any detectable phenotype, including but not limited to susceptibility to a disease or response to a treatment. The method involves obtaining data on the frequency of the genotype(s), haplotype(s), or haplotype pair(s) of interest in a reference population as well as in a population exhibiting the trait. Frequency data for one or both of the reference and trait populations may be obtained by genotyping or haplotyping each individual in the populations using one of the methods described above. The haplotypes for the trait population may be determined directly or, alternatively, by the predictive genotype to haplotype approach described above. In another embodiment, the frequency data for the reference and/or trait populations is obtained by accessing previously determined frequency data, which may be in written or electronic form. For example, the frequency data may be present in a database that is accessible by a computer. Once the frequency data is obtained, the frequencies of the genotype(s),

haplotype(s), or haplotype pair(s) of interest in the reference and trait populations are compared. In a preferred embodiment, the frequencies of all genotypes, haplotypes, and/or haplotype pairs observed in the populations are compared. If a particular EDG4 genotype, haplotype, or haplotype pair is more frequent in the trait population than in the reference population at a statistically significant amount, then the trait is predicted to be associated with that EDG4 genotype, haplotype or haplotype pair. Preferably, the EDG4 genotype, haplotype, or haplotype pair being compared in the trait and reference populations is selected from the full-genotypes and full-haplotypes shown in Tables 3 and 4, or from sub-genotypes and sub-haplotypes derived from these genotypes and haplotypes.

In a preferred embodiment of the method, the trait of interest is a clinical response exhibited by a patient to some therapeutic treatment, for example, response to a drug targeting EDG4 or response to a therapeutic treatment for a medical condition. As used herein, "medical condition" includes but is not limited to any condition or disease manifested as one or more physical and/or psychological symptoms for which treatment is desirable, and includes previously and newly identified diseases and other disorders. As used herein the term "clinical response" means any or all of the following: a quantitative measure of the response, no response, and adverse response (i.e., side effects).

In order to deduce a correlation between clinical response to a treatment and an EDG4 genotype, haplotype, or haplotype pair, it is necessary to obtain data on the clinical responses exhibited by a population of individuals who received the treatment, hereinafter the "clinical population". This clinical data may be obtained by analyzing the results of a clinical trial that has already been run and/or the clinical data may be obtained by designing and carrying out one or more new clinical trials. As used herein, the term "clinical trial" means any research study designed to collect clinical data on responses to a particular treatment, and includes but is not limited to phase I, phase II and phase III clinical trials. Standard methods are used to define the patient population and to enroll subjects.

It is preferred that the individuals included in the clinical population have been graded for the existence of the medical condition of interest. This is important in cases where the symptom(s) being presented by the patients can be caused by more than one underlying condition, and where treatment of the underlying conditions are not the same. An example of this would be where patients experience breathing difficulties that are due to either asthma or respiratory infections. If both sets were treated with an asthma medication, there would be a spurious group of apparent non-responders that did not actually have asthma. These people would affect the ability to detect any correlation between haplotype and treatment outcome. This grading of potential patients could employ a standard physical exam or one or more lab tests. Alternatively, grading of patients could use haplotyping for situations where there is a strong correlation between haplotype pair and disease susceptibility or severity.

The therapeutic treatment of interest is administered to each individual in the trial population and each individual's response to the treatment is measured using one or more predetermined criteria. It is contemplated that in many cases, the trial population will exhibit a range of responses and that the investigator will choose the number of responder groups (e.g., low, medium, high) made up by the

various responses. In addition, the EDG4 gene for each individual in the trial population is genotyped and/or haplotyped, which may be done before or after administering the treatment.

After both the clinical and polymorphism data have been obtained, correlations between individual response and EDG4 genotype or haplotype content are created. Correlations may be produced in several ways. In one method, individuals are grouped by their EDG4 genotype or haplotype (or haplotype pair) (also referred to as a polymorphism group), and then the averages and standard deviations of clinical responses exhibited by the members of each polymorphism group are calculated.

These results are then analyzed to determine if any observed variation in clinical response between polymorphism groups is statistically significant. Statistical analysis methods which may be used are described in L.D. Fisher and G. vanBelle, "Biostatistics: A Methodology for the Health Sciences", Wiley-Interscience (New York) 1993. This analysis may also include a regression calculation of which polymorphic sites in the EDG4 gene give the most significant contribution to the differences in phenotype. One regression model useful in the invention is described in PCT Application Serial No. PCT/US00/17540, entitled "Methods for Obtaining and Using Haplotype Data".

A second method for finding correlations between EDG4 haplotype content and clinical responses uses predictive models based on error-minimizing optimization algorithms. One of many possible optimization algorithms is a genetic algorithm (R. Judson, "Genetic Algorithms and Their Uses in Chemistry" in Reviews in Computational Chemistry, Vol. 10, pp. 1-73, K. B. Lipkowitz and D. B. Boyd, eds. (VCH Publishers, New York, 1997). Simulated annealing (Press et al., "Numerical Recipes in C: The Art of Scientific Computing", Cambridge University Press (Cambridge) 1992, Ch. 10), neural networks (E. Rich and K. Knight, "Artificial Intelligence", 2<sup>nd</sup> Edition (McGraw-Hill, New York, 1991, Ch. 18), standard gradient descent methods (Press et al., *supra*, Ch. 10), or other global or local optimization approaches (see discussion in Judson, *supra*) could also be used. Preferably, the correlation is found using a genetic algorithm approach as described in PCT Application Serial No. PCT/US00/17540.

Correlations may also be analyzed using analysis of variation (ANOVA) techniques to determine how much of the variation in the clinical data is explained by different subsets of the polymorphic sites in the EDG4 gene. As described in PCT Application Serial No. PCT/US00/17540, ANOVA is used to test hypotheses about whether a response variable is caused by or correlated with one or more traits or variables that can be measured (Fisher and vanBelle, *supra*, Ch. 10).

From the analyses described above, a mathematical model may be readily constructed by the skilled artisan that predicts clinical response as a function of EDG4 genotype or haplotype content. Preferably, the model is validated in one or more follow-up clinical trials designed to test the model.

The identification of an association between a clinical response and a genotype or haplotype (or haplotype pair) for the EDG4 gene may be the basis for designing a diagnostic method to determine those individuals who will or will not respond to the treatment, or alternatively, will respond at a lower

level and thus may require more treatment, i.e., a greater dose of a drug. The diagnostic method may take one of several forms: for example, a direct DNA test (i.e., genotyping or haplotyping one or more of the polymorphic sites in the EDG4 gene), a serological test, or a physical exam measurement. The only requirement is that there be a good correlation between the diagnostic test results and the underlying EDG4 genotype or haplotype that is in turn correlated with the clinical response. In a preferred embodiment, this diagnostic method uses the predictive haplotyping method described above.

In another embodiment, the invention provides an isolated polynucleotide comprising a polymorphic variant of the EDG4 gene or a fragment of the gene which contains at least one of the novel polymorphic sites described herein. The nucleotide sequence of a variant EDG4 gene is identical to the reference genomic sequence for those portions of the gene examined, as described in the Examples below, except that it comprises a different nucleotide at one or more of the novel polymorphic sites PS1, PS2, PS3, PS4, PS5, PS6, PS7 and PS8. Similarly, the nucleotide sequence of a variant fragment of the EDG4 gene is identical to the corresponding portion of the reference sequence except for having a different nucleotide at one or more of the novel polymorphic sites described herein. Thus, the invention specifically does not include polynucleotides comprising a nucleotide sequence identical to the reference sequence of the EDG4 gene, which is defined by haplotype 5, (or other reported EDG4 sequences) or to portions of the reference sequence (or other reported EDG4 sequences), except for genotyping oligonucleotides as described above.

The location of a polymorphism in a variant gene or fragment is identified by aligning its sequence against SEQ ID NO:1. The polymorphism is selected from the group consisting of cytosine at PS1, thymine at PS2, guanine at PS3, adenine at PS4, adenine at PS5, thymine at PS6, thymine at PS7 and thymine at PS8. In a preferred embodiment, the polymorphic variant comprises a naturally-occurring isogene of the EDG4 gene which is defined by any one of haplotypes 1- 4 and 6 - 8 shown in Table 4 below.

Polymorphic variants of the invention may be prepared by isolating a clone containing the EDG4 gene from a human genomic library. The clone may be sequenced to determine the identity of the nucleotides at the novel polymorphic sites described herein. Any particular variant claimed herein could be prepared from this clone by performing *in vitro* mutagenesis using procedures well-known in the art.

EDG4 isogenes may be isolated using any method that allows separation of the two "copies" of the EDG4 gene present in an individual, which, as readily understood by the skilled artisan, may be the same allele or different alleles. Separation methods include targeted *in vivo* cloning (TIVC) in yeast as described in WO 98/01573, U.S. Patent No. 5,866,404, and U.S. Patent No. 5,972,614. Another method, which is described in U.S. Patent No. 5,972,614, uses an allele specific oligonucleotide in combination with primer extension and exonuclease degradation to generate hemizygous DNA targets. Yet other methods are single molecule dilution (SMD) as described in Ruaño et al., *Proc. Natl. Acad. Sci.* 87:6296-6300, 1990; and allele specific PCR (Ruaño et al., 1989, *supra*; Ruaño et al., 1991, *supra*;

Michalatos-Beloin et al., *supra*).

The invention also provides EDG4 genome anthologies, which are collections of EDG4 isogenes found in a given population. The population may be any group of at least two individuals, including but not limited to a reference population, a population group, a family population, a clinical population, and a same sex population. An EDG4 genome anthology may comprise individual EDG4 isogenes stored in separate containers such as microtest tubes, separate wells of a microtitre plate and the like. Alternatively, two or more groups of the EDG4 isogenes in the anthology may be stored in separate containers. Individual isogenes or groups of isogenes in a genome anthology may be stored in any convenient and stable form, including but not limited to in buffered solutions, as DNA precipitates, freeze-dried preparations and the like. A preferred EDG4 genome anthology of the invention comprises a set of isogenes defined by the haplotypes shown in Table 4 below.

An isolated polynucleotide containing a polymorphic variant nucleotide sequence of the invention may be operably linked to one or more expression regulatory elements in a recombinant expression vector capable of being propagated and expressing the encoded EDG4 protein in a prokaryotic or a eukaryotic host cell. Examples of expression regulatory elements which may be used include, but are not limited to, the lac system, operator and promoter regions of phage lambda, yeast promoters, and promoters derived from vaccinia virus, adenovirus, retroviruses, or SV40. Other regulatory elements include, but are not limited to, appropriate leader sequences, termination codons, polyadenylation signals, and other sequences required for the appropriate transcription and subsequent translation of the nucleic acid sequence in a given host cell. Of course, the correct combinations of expression regulatory elements will depend on the host system used. In addition, it is understood that the expression vector contains any additional elements necessary for its transfer to and subsequent replication in the host cell. Examples of such elements include, but are not limited to, origins of replication and selectable markers. Such expression vectors are commercially available or are readily constructed using methods known to those in the art (e.g., F. Ausubel et al., 1987, in "Current Protocols in Molecular Biology", John Wiley and Sons, New York, New York). Host cells which may be used to express the variant EDG4 sequences of the invention include, but are not limited to, eukaryotic and mammalian cells, such as animal, plant, insect and yeast cells, and prokaryotic cells, such as *E. coli*, or algal cells as known in the art. The recombinant expression vector may be introduced into the host cell using any method known to those in the art including, but not limited to, microinjection, electroporation, particle bombardment, transduction, and transfection using DEAE-dextran, lipofection, or calcium phosphate (see e.g., Sambrook et al. (1989) in "Molecular Cloning. A Laboratory Manual", Cold Spring Harbor Press, Plainview, New York). In a preferred aspect, eukaryotic expression vectors that function in eukaryotic cells, and preferably mammalian cells, are used. Non-limiting examples of such vectors include vaccinia virus vectors, adenovirus vectors, herpes virus vectors, and baculovirus transfer vectors. Preferred eukaryotic cell lines include COS cells, CHO cells, HeLa cells, NIH/3T3 cells, and embryonic stem cells (Thomson, J. A. et al., 1998 *Science* 282:1145-1147). Particularly



preferred host cells are mammalian cells.

As will be readily recognized by the skilled artisan, expression of polymorphic variants of the EDG4 gene will produce EDG4 mRNAs varying from each other at any polymorphic site retained in the spliced and processed mRNA molecules. These mRNAs can be used for the preparation of an EDG4 cDNA comprising a nucleotide sequence which is a polymorphic variant of the EDG4 reference coding sequence shown in Figure 2. Thus, the invention also provides EDG4 mRNAs and corresponding cDNAs which comprise a nucleotide sequence that is identical to SEQ ID NO:2 (Fig. 2), or its corresponding RNA sequence, except for having one or more polymorphisms selected from the group consisting of guanine at a position corresponding to nucleotide 125, adenine at a position corresponding to nucleotide 408 and adenine at a position corresponding to nucleotide 414. A particularly preferred polymorphic cDNA variant comprises the coding sequence of an EDG4 isogene defined by haplotypes 1c-4c and 7c. Fragments of these variant mRNAs and cDNAs are included in the scope of the invention, provided they contain the novel polymorphisms described herein. The invention specifically excludes polynucleotides identical to previously identified and characterized EDG4 cDNAs and fragments thereof. Polynucleotides comprising a variant RNA or DNA sequence may be isolated from a biological sample using well-known molecular biological procedures or may be chemically synthesized.

As used herein, a polymorphic variant of an EDG4 gene fragment comprises at least one novel polymorphism identified herein and has a length of at least 10 nucleotides and may range up to the full length of the gene. Preferably, such fragments are between 100 and 3000 nucleotides in length, and more preferably between 200 and 2000 nucleotides in length, and most preferably between 500 and 1000 nucleotides in length.

In describing the EDG4 polymorphic sites identified herein, reference is made to the sense strand of the gene for convenience. However, as recognized by the skilled artisan, nucleic acid molecules containing the EDG4 gene may be complementary double stranded molecules and thus reference to a particular site on the sense strand refers as well to the corresponding site on the complementary antisense strand. Thus, reference may be made to the same polymorphic site on either strand and an oligonucleotide may be designed to hybridize specifically to either strand at a target region containing the polymorphic site. Thus, the invention also includes single-stranded polynucleotides which are complementary to the sense strand of the EDG4 genomic variants described herein.

Polynucleotides comprising a polymorphic gene variant or fragment may be useful for therapeutic purposes. For example, where a patient could benefit from expression, or increased expression, of a particular EDG4 protein isoform, an expression vector encoding the isoform may be administered to the patient. The patient may be one who lacks the EDG4 isogene encoding that isoform or may already have at least one copy of that isogene.

In other situations, it may be desirable to decrease or block expression of a particular EDG4

isogene. Expression of an EDG4 isogene may be turned off by transforming a targeted organ, tissue or cell population with an expression vector that expresses high levels of untranslatable mRNA for the isogene. Alternatively, oligonucleotides directed against the regulatory regions (e.g., promoter, introns, enhancers, 3' untranslated region) of the isogene may block transcription. Oligonucleotides targeting the transcription initiation site, e.g., between positions -10 and +10 from the start site are preferred. Similarly, inhibition of transcription can be achieved using oligonucleotides that base-pair with region(s) of the isogene DNA to form triplex DNA (see e.g., Gee et al. in Huber, B.E. and B.I. Carr, Molecular and Immunologic Approaches, Futura Publishing Co., Mt. Kisco, N.Y., 1994). Antisense oligonucleotides may also be designed to block translation of EDG4 mRNA transcribed from a particular isogene. It is also contemplated that ribozymes may be designed that can catalyze the specific cleavage of EDG4 mRNA transcribed from a particular isogene.

The oligonucleotides may be delivered to a target cell or tissue by expression from a vector introduced into the cell or tissue *in vivo* or *ex vivo*. Alternatively, the oligonucleotides may be formulated as a pharmaceutical composition for administration to the patient. Oligoribonucleotides and/or oligodeoxynucleotides intended for use as antisense oligonucleotides may be modified to increase stability and half-life. Possible modifications include, but are not limited to phosphorothioate or 2' O-methyl linkages, and the inclusion of nontraditional bases such as inosine and queosine, as well as acetyl-, methyl-, thio-, and similarly modified forms of adenine, cytosine, guanine, thymine, and uracil which are not as easily recognized by endogenous nucleases.

The invention also provides an isolated polypeptide comprising a polymorphic variant of the reference EDG4 amino acid sequence shown in Figure 3. The location of a variant amino acid in an EDG4 polypeptide or fragment of the invention is identified by aligning its sequence against SEQ ID NO:3 (Fig. 3). An EDG4 protein variant of the invention comprises an amino acid sequence identical to SEQ ID NO:3 except for having glycine at a position corresponding to amino acid position 42. The invention specifically excludes amino acid sequences identical to those previously identified for EDG4, including SEQ ID NO:3, and previously described fragments thereof. EDG4 protein variants included within the invention comprise all amino acid sequences based on SEQ ID NO:3 and having glycine at a position corresponding to amino acid position 42.

The invention also includes EDG4 peptide variants, which are any fragments of an EDG4 protein variant that contain glycine at a position corresponding to amino acid position 42. An EDG4 peptide variant is at least 6 amino acids in length and is preferably any number between 6 and 30 amino acids long, more preferably between 10 and 25, and most preferably between 15 and 20 amino acids long. Such EDG4 peptide variants may be useful as antigens to generate antibodies specific for one of the above EDG4 isoforms. In addition, the EDG4 peptide variants may be useful in drug screening assays.

An EDG4 variant protein or peptide of the invention may be prepared by chemical synthesis or by expressing one of the variant EDG4 genomic and cDNA sequences as described above.

Alternatively, the EDG4 protein variant may be isolated from a biological sample of an individual having an EDG4 isogene which encodes the variant protein. Where the sample contains two different EDG4 isoforms (i.e., the individual has different EDG4 isogenes), a particular EDG4 isoform of the invention can be isolated by immunoaffinity chromatography using an antibody which specifically binds to that particular EDG4 isoform but does not bind to the other EDG4 isoform.

The expressed or isolated EDG4 protein may be detected by methods known in the art, including Coomassie blue staining, silver staining, and Western blot analysis using antibodies specific for the isoform of the EDG4 protein as discussed further below. EDG4 variant proteins can be purified by standard protein purification procedures known in the art, including differential precipitation, molecular sieve chromatography, ion-exchange chromatography, isoelectric focusing, gel electrophoresis, affinity and immunoaffinity chromatography and the like. (Ausubel et. al., 1987, In Current Protocols in Molecular Biology John Wiley and Sons, New York, New York). In the case of immunoaffinity chromatography, antibodies specific for a particular polymorphic variant may be used.

A polymorphic variant EDG4 gene of the invention may also be fused in frame with a heterologous sequence to encode a chimeric EDG4 protein. The non-EDG4 portion of the chimeric protein may be recognized by a commercially available antibody. In addition, the chimeric protein may also be engineered to contain a cleavage site located between the EDG4 and non-EDG4 portions so that the EDG4 protein may be cleaved and purified away from the non-EDG4 portion.

An additional embodiment of the invention relates to using a novel EDG4 protein isoform in any of a variety of drug screening assays. Such screening assays may be performed to identify agents that bind specifically to all known EDG4 protein isoforms or to only a subset of one or more of these isoforms. The agents may be from chemical compound libraries, peptide libraries and the like. The EDG4 protein or peptide variant may be free in solution or affixed to a solid support. In one embodiment, high throughput screening of compounds for binding to an EDG4 variant may be accomplished using the method described in PCT application WO84/03565, in which large numbers of test compounds are synthesized on a solid substrate, such as plastic pins or some other surface, contacted with the EDG4 protein(s) of interest and then washed. Bound EDG4 protein(s) are then detected using methods well-known in the art.

In another embodiment, a novel EDG4 protein isoform may be used in assays to measure the binding affinities of one or more candidate drugs targeting the EDG4 protein.

In yet another embodiment, when a particular EDG4 haplotype or group of EDG4 haplotypes encodes an EDG4 protein variant with an amino acid sequence distinct from that of EDG4 protein isoforms encoded by other EDG4 haplotypes, then detection of that particular EDG4 haplotype or group of EDG4 haplotypes may be accomplished by detecting expression of the encoded EDG4 protein variant using any of the methods described herein or otherwise commonly known to the skilled artisan.

In another embodiment, the invention provides antibodies specific for and immunoreactive with one or more of the novel EDG4 variant proteins described herein. The antibodies may be either

monoclonal or polyclonal in origin. The EDG4 protein or peptide variant used to generate the antibodies may be from natural or recombinant sources or produced by chemical synthesis using synthesis techniques known in the art. If the EDG4 protein variant is of insufficient size to be antigenic, it may be conjugated, complexed, or otherwise covalently linked to a carrier molecule to enhance the antigenicity of the peptide. Examples of carrier molecules, include, but are not limited to, albumins (e.g., human, bovine, fish, ovine), and keyhole limpet hemocyanin (Basic and Clinical Immunology, 1991, Eds. D.P. Stites, and A.I. Terr, Appleton and Lange, Norwalk Connecticut, San Mateo, California).

In one embodiment, an antibody specifically immunoreactive with one of the novel protein isoforms described herein is administered to an individual to neutralize activity of the EDG4 isoform expressed by that individual. The antibody may be formulated as a pharmaceutical composition which includes a pharmaceutically acceptable carrier.

Antibodies specific for and immunoreactive with one of the novel protein isoforms described herein may be used to immunoprecipitate the EDG4 protein variant from solution as well as react with EDG4 protein isoforms on Western or immunoblots of polyacrylamide gels on membrane supports or substrates. In another preferred embodiment, the antibodies will detect EDG4 protein isoforms in paraffin or frozen tissue sections, or in cells which have been fixed or unfixed and prepared on slides, coverslips, or the like, for use in immunocytochemical, immunohistochemical, and immunofluorescence techniques.

In another embodiment, an antibody specifically immunoreactive with one of the novel EDG4 protein variants described herein is used in immunoassays to detect this variant in biological samples. In this method, an antibody of the present invention is contacted with a biological sample and the formation of a complex between the EDG4 protein variant and the antibody is detected. As described, suitable immunoassays include radioimmunoassay, Western blot assay, immunofluorescent assay, enzyme linked immunoassay (ELISA), chemiluminescent assay, immunohistochemical assay, immunocytochemical assay, and the like (see, e.g., Principles and Practice of Immunoassay, 1991, Eds. Christopher P. Price and David J. Neoman, Stockton Press, New York, New York; Current Protocols in Molecular Biology, 1987, Eds. Ausubel et al., John Wiley and Sons, New York, New York). Standard techniques known in the art for ELISA are described in Methods in Immunodiagnosis, 2nd Ed., Eds. Rose and Bigazzi, John Wiley and Sons, New York 1980; and Campbell et al., 1984, Methods in Immunology, W.A. Benjamin, Inc.). Such assays may be direct, indirect, competitive, or noncompetitive as described in the art (see, e.g., Principles and Practice of Immunoassay, 1991, Eds. Christopher P. Price and David J. Neoman, Stockton Pres, NY, NY; and Oellirich, M., 1984, J. Clin. Chem. Clin. Biochem., 22:895-904). Proteins may be isolated from test specimens and biological samples by conventional methods, as described in Current Protocols in Molecular Biology, supra.

Exemplary antibody molecules for use in the detection and therapy methods of the present invention are intact immunoglobulin molecules, substantially intact immunoglobulin molecules, or

those portions of immunoglobulin molecules that contain the antigen binding site. Polyclonal or monoclonal antibodies may be produced by methods conventionally known in the art (e.g., Kohler and Milstein, 1975, Nature, 256:495-497; Campbell Monoclonal Antibody Technology, the Production and Characterization of Rodent and Human Hybridomas, 1985, In: Laboratory Techniques in Biochemistry and Molecular Biology, Eds. Burdon et al., Volume 13, Elsevier Science Publishers, Amsterdam). The antibodies or antigen binding fragments thereof may also be produced by genetic engineering. The technology for expression of both heavy and light chain genes in E. coli is the subject of PCT patent applications, publication number WO 901443, WO 901443 and WO 9014424 and in Huse et al., 1989, Science, 246:1275-1281. The antibodies may also be humanized (e.g., Queen, C. et al. 1989 Proc. Natl. Acad. Sci. USA 86:10029).

Effect(s) of the polymorphisms identified herein on expression of EDG4 may be investigated by preparing recombinant cells and/or nonhuman recombinant organisms, preferably recombinant animals, containing a polymorphic variant of the EDG4 gene. As used herein, "expression" includes but is not limited to one or more of the following: transcription of the gene into precursor mRNA; splicing and other processing of the precursor mRNA to produce mature mRNA; mRNA stability; translation of the mature mRNA into EDG4 protein (including codon usage and tRNA availability); and glycosylation and/or other modifications of the translation product, if required for proper expression and function.

To prepare a recombinant cell of the invention, the desired EDG4 isogene may be introduced into the cell in a vector such that the isogene remains extrachromosomal. In such a situation, the gene will be expressed by the cell from the extrachromosomal location. In a preferred embodiment, the EDG4 isogene is introduced into a cell in such a way that it recombines with the endogenous EDG4 gene present in the cell. Such recombination requires the occurrence of a double recombination event, thereby resulting in the desired EDG4 gene polymorphism. Vectors for the introduction of genes both for recombination and for extrachromosomal maintenance are known in the art, and any suitable vector or vector construct may be used in the invention. Methods such as electroporation, particle bombardment, calcium phosphate co-precipitation and viral transduction for introducing DNA into cells are known in the art; therefore, the choice of method may lie with the competence and preference of the skilled practitioner. Examples of cells into which the EDG4 isogene may be introduced include, but are not limited to, continuous culture cells, such as COS, NIH/3T3, and primary or culture cells of the relevant tissue type, i.e., they express the EDG4 isogene. Such recombinant cells can be used to compare the biological activities of the different protein variants.

Recombinant nonhuman organisms, i.e., transgenic animals, expressing a variant EDG4 gene are prepared using standard procedures known in the art. Preferably, a construct comprising the variant gene is introduced into a nonhuman animal or an ancestor of the animal at an embryonic stage, i.e., the one-cell stage, or generally not later than about the eight-cell stage. Transgenic animals carrying the constructs of the invention can be made by several methods known to those having skill in the art. One

method involves transfecting into the embryo a retrovirus constructed to contain one or more insulator elements, a gene or genes of interest, and other components known to those skilled in the art to provide a complete shuttle vector harboring the insulated gene(s) as a transgene, see e.g., U.S. Patent No.

5,610,053. Another method involves directly injecting a transgene into the embryo. A third method involves the use of embryonic stem cells. Examples of animals into which the EDG4 isogenes may be introduced include, but are not limited to, mice, rats, other rodents, and nonhuman primates (see "The Introduction of Foreign Genes into Mice" and the cited references therein, In: Recombinant DNA, Eds. J.D. Watson, M. Gilman, J. Witkowski, and M. Zoller; W.H. Freeman and Company, New York, pages 254-272). Transgenic animals stably expressing a human EDG4 isogene and producing human EDG4 protein can be used as biological models for studying diseases related to abnormal EDG4 expression and/or activity, and for screening and assaying various candidate drugs, compounds, and treatment regimens to reduce the symptoms or effects of these diseases.

An additional embodiment of the invention relates to pharmaceutical compositions for treating disorders affected by expression or function of a novel EDG4 isogene described herein. The pharmaceutical composition may comprise any of the following active ingredients: a polynucleotide comprising one of these novel EDG4 isogenes; an antisense oligonucleotide directed against one of the novel EDG4 isogenes, a polynucleotide encoding such an antisense oligonucleotide, or another compound which inhibits expression of a novel EDG4 isogene described herein. Preferably, the composition contains the active ingredient in a therapeutically effective amount. By therapeutically effective amount is meant that one or more of the symptoms relating to disorders affected by expression or function of a novel EDG4 isogene is reduced and/or eliminated. The composition also comprises a pharmaceutically acceptable carrier, examples of which include, but are not limited to, saline, buffered saline, dextrose, and water. Those skilled in the art may employ a formulation most suitable for the active ingredient, whether it is a polynucleotide, oligonucleotide, protein, peptide or small molecule antagonist. The pharmaceutical composition may be administered alone or in combination with at least one other agent, such as a stabilizing compound. Administration of the pharmaceutical composition may be by any number of routes including, but not limited to oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, intradermal, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal. Further details on techniques for formulation and administration may be found in the latest edition of Remington's Pharmaceutical Sciences (Maack Publishing Co., Easton, PA).

For any composition, determination of the therapeutically effective dose of active ingredient and/or the appropriate route of administration is well within the capability of those skilled in the art. For example, the dose can be estimated initially either in cell culture assays or in animal models. The animal model may also be used to determine the appropriate concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans. The exact dosage will be determined by the practitioner, in light of factors

relating to the patient requiring treatment, including but not limited to severity of the disease state, general health, age, weight and gender of the patient, diet, time and frequency of administration, other drugs being taken by the patient, and tolerance/response to the treatment.

Any or all analytical and mathematical operations involved in practicing the methods of the present invention may be implemented by a computer. In addition, the computer may execute a program that generates views (or screens) displayed on a display device and with which the user can interact to view and analyze large amounts of information relating to the EDG4 gene and its genomic variation, including chromosome location, gene structure, and gene family, gene expression data, polymorphism data, genetic sequence data, and clinical data population data (e.g., data on ethnogeographic origin, clinical responses, genotypes, and haplotypes for one or more populations). The EDG4 polymorphism data described herein may be stored as part of a relational database (e.g., an instance of an Oracle database or a set of ASCII flat files). These polymorphism data may be stored on the computer's hard drive or may, for example, be stored on a CD-ROM or on one or more other storage devices accessible by the computer. For example, the data may be stored on one or more databases in communication with the computer via a network.

Preferred embodiments of the invention are described in the following examples. Other embodiments within the scope of the claims herein will be apparent to one skilled in the art from consideration of the specification or practice of the invention as disclosed herein. It is intended that the specification, together with the examples, be considered exemplary only, with the scope and spirit of the invention being indicated by the claims which follow the examples.

## EXAMPLES

The Examples herein are meant to exemplify the various aspects of carrying out the invention and are not intended to limit the scope of the invention in any way. The Examples do not include detailed descriptions for conventional methods employed, such as in the performance of genomic DNA isolation, PCR and sequencing procedures. Such methods are well-known to those skilled in the art and are described in numerous publications, for example, Sambrook, Fritsch, and Maniatis, "Molecular Cloning: A Laboratory Manual", 2<sup>nd</sup> Edition, Cold Spring Harbor Laboratory Press, USA, (1989).

### EXAMPLE 1

This example illustrates examination of various regions of the EDG4 gene for polymorphic sites.

#### Amplification of Target Regions

The following target regions of the EDG4 gene were amplified using PCR primer pairs. The primers used for each region are represented below by providing the nucleotide positions of their initial and final nucleotides, which correspond to positions in Figure 1.

PCR Primer Pairs

	Fragment No.	Forward Primer	Reverse Primer	PCR Product
	Fragment 1	30-52	complement of 617-595	588 nt
5	Fragment 2	240-262	complement of 859-839	620 nt
	Fragment 3	408-429	complement of 1047-1027	640 nt
	Fragment 4	692-714	complement of 1325-1302	634 nt
	Fragment 5	2883-2906	complement of 3426-3404	544 nt
10	Fragment 6	3145-3165	complement of 3624-3604	480 nt

These primer pairs were used in PCR reactions containing genomic DNA isolated from immortalized cell lines for each member of the Index Repository. The PCR reactions were carried out under the following conditions:

15	Reaction volume	= 10 $\mu$ l
	10 x Advantage 2 Polymerase reaction buffer (Clontech)	= 1 $\mu$ l
	100 ng of human genomic DNA	= 1 $\mu$ l
	10 mM dNTP	= 0.4 $\mu$ l
	Advantage 2 Polymerase enzyme mix (Clontech)	= 0.2 $\mu$ l
20	Forward Primer (10 $\mu$ M)	= 0.4 $\mu$ l
	Reverse Primer (10 $\mu$ M)	= 0.4 $\mu$ l
	Water	= 6.6 $\mu$ l

Amplification profile:

25	97°C - 2 min.	1 cycle
	97°C - 15 sec.	} 10 cycles
	70°C - 45 sec.	
	72°C - 45 sec.	
30		
	97°C - 15 sec.	} 35 cycles
	64°C - 45 sec.	
	72°C - 45 sec.	
35		

Sequencing of PCR Products

The PCR products were purified using a Whatman/Polyfiltronics 100  $\mu$ l 384 well unfilter plate essentially according to the manufacturers protocol. The purified DNA was eluted in 50  $\mu$ l of distilled water. Sequencing reactions were set up using Applied Biosystems Big Dye Terminator chemistry essentially according to the manufacturers protocol. The purified PCR products were sequenced in both directions using the primer sets described previously or those represented below by the nucleotide positions of their initial and final nucleotides, which correspond to positions in Figure 1. Reaction products were purified by isopropanol precipitation, and run on an Applied Biosystems 3700 DNA Analyzer.



Sequencing Primer Pairs

Fragment No.	Forward Primer	Reverse Primer
Fragment 1	98-117	complement of 554-534
Fragment 2	265-284	complement of 750-731
Fragment 3	535-554	complement of 1017-998
Fragment 4	791-809	complement of 1273-1253
Fragment 5	2994-3014	complement of 3365-3346
Fragment 6	3171-3191	complement of 3549-3530

Analysis of Sequences for Polymorphic Sites

Sequence information for a minimum of 80 humans was analyzed for the presence of polymorphisms using the Polyphred program (Nickerson et al., *Nucleic Acids Res.* 14:2745-2751, 1997). The presence of a polymorphism was confirmed on both strands. The polymorphisms and their positions in the EDG4 gene are listed in Table 2 below.

Table 2. Polymorphic Sites Identified in the EDG4 Gene

Polymorphic Site Number	PolyId <sup>a</sup>	Nucleotide Position	Reference Allele	Variant Allele	CDS Position	Variant AA
PS1	1721433	216	T	C		
PS2	1721431	395	C	T		
PS3	1721429	523	T	G	125	V42G
PS4	1721423	806	G	A	408	V136V
PS5	1721421	812	G	A	414	L138L
PS6	1721411	1291	G	T		
PS7	1721405	3099	G	T		
PS8	1721399	3449	C	T		

<sup>a</sup>PolyId is a unique identifier assigned to each PS by Genaisance Pharmaceuticals, Inc.

## EXAMPLE 2

This example illustrates analysis of the EDG4 polymorphisms identified in the Index Repository for human genotypes and haplotypes.

The different genotypes containing these polymorphisms that were observed in the reference population are shown in Table 3 below, with the haplotype pair indicating the combination of haplotypes determined for the individual using the haplotype derivation protocol described below. In Table 3, homozygous positions are indicated by one nucleotide and heterozygous positions are indicated by two nucleotides. Missing nucleotides in any given genotype in Table 3 were inferred based on linkage disequilibrium and/or Mendelian inheritance.

Table 3. Genotypes and Haplotype Pairs Observed for EDG4 Gene

Genotype		Polymorphic Sites								HAP Pair	
Number		PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8		
5	1	T	C	T	G	G	G	G	C	5	5
	2	T	C	T	A	G	G	G	C	4	4
	3	T/C	C	T	G	G/A	G/T	G	C	5	1
	4	T	C	T/G	G	G	G	G	C	5	3
	5	T	C/T	T	G/A	G	G	G/T	C	5	7
10	6	T	C	T	G/A	G	G	G	C	5	4
	7	T	C/T	T	G	G	G	G	C	5	8
	8	T	C	T	G	G	G	G	C/T	5	6
	9	T	C	T/G	G/A	G	G	G	C	5	2

The haplotype pairs shown in Table 3 were estimated from the unphased genotypes using a computer-implemented extension of Clark's algorithm (Clark, A.G. 1990 *Mol Bio Evol* 7, 111-122) for assigning haplotypes to unrelated individuals in a population sample, as described in U.S. Provisional Application Serial No. 60/198,340 entitled "A Method and System for Determining Haplotypes from a Collection of Polymorphisms" and the corresponding International Application, PCT/US01/12831. In this method, haplotypes are assigned directly from individuals who are homozygous at all sites or heterozygous at no more than one of the variable sites. This list of haplotypes is then used to deconvolute the unphased genotypes in the remaining (multiply heterozygous) individuals. In our analysis, the list of haplotypes was augmented with haplotypes obtained from two families (one three-generation Caucasian family and one two-generation African-American family).

By following this protocol, it was determined that the Index Repository examined herein and, by extension, the general population contains the 8 human EDG4 haplotypes shown in Table 4 below.

An EDG4 isogene defined by a full-haplotype shown in Table 4 below comprises the regions of the SEQ ID NOS indicated in Table 4, with their corresponding set of polymorphic locations and identities, which are also set forth in Table 4.

Table 4. Haplotypes Identified in the EDG4 Gene

	Haplotype Number <sup>a</sup>								PS	PS	SEQ ID	Region
	1	2	3	4	5	6	7	8	No. <sup>b</sup>	Pos. <sup>c</sup>	No. <sup>d</sup>	Examined <sup>e</sup>
5	C	T	T	T	T	T	T	T	1	216/30	1/44	30-1325
	C	C	C	C	C	C	T	T	2	395/150	1/44	30-1325
	T	G	G	T	T	T	T	T	3	523/270	1/44	30-1325
	G	A	G	A	G	G	A	G	4	806/390	1/44	30-1325
	A	G	G	G	G	G	G	G	5	812/510	1/44	30-1325
10	T	G	G	G	G	G	G	G	6	1291/630	1/44	30-1325
	G	G	G	G	G	G	T	G	7	3099/750	1/44	2883-3624
	C	C	C	C	C	T	C	C	8	3449/870	1/44	2883-3624

<sup>a</sup>Alleles for EDG4 haplotypes are presented 5' to 3' in each column;

<sup>b</sup>PS = polymorphic site;

<sup>c</sup>Position of PS within the indicated SEQ ID NO, with the 1<sup>st</sup> position number referring to the first SEQ ID NO and the 2<sup>nd</sup> position number referring to the 2<sup>nd</sup> SEQ ID NO;

<sup>d</sup>1<sup>st</sup> SEQ ID NO refers to Figure 1, with the two alternative allelic variants of each polymorphic site indicated by the appropriate nucleotide symbol; 2<sup>nd</sup> SEQ ID NO is a modified version of the 1<sup>st</sup> SEQ ID NO that comprises the context sequence of each polymorphic site, PS1-PS8, to facilitate electronic searching of the haplotypes;

<sup>e</sup>Region examined represents the nucleotide positions defining the start and stop positions within the 1<sup>st</sup> SEQ ID NO of the sequenced region.

SEQ ID NO:1 refers to Figure 1, with the two alternative allelic variants of each polymorphic site indicated by the appropriate nucleotide symbol. SEQ ID NO:44 is a modified version of SEQ ID NO:1 that shows the context sequence of each of PS1-PS8 in a uniform format to facilitate electronic searching of the EDG4 haplotypes. For each polymorphic site, SEQ ID NO:44 contains a block of 60 bases of the nucleotide sequence encompassing the centrally-located polymorphic site at the 30<sup>th</sup> position, followed by 60 bases of unspecified sequence to represent that each polymorphic site is separated by genomic sequence whose composition is defined elsewhere herein.

Table 5 below shows the percent of chromosomes characterized by a given EDG4 haplotype for all unrelated individuals in the Index Repository for which haplotype data was obtained. The percent of these unrelated individuals who have a given EDG4 haplotype pair is shown in Table 6. In Tables 5 and 6, the "Total" column shows this frequency data for all of these unrelated individuals, while the other columns show the frequency data for these unrelated individuals categorized according to their self-identified ethnogeographic origin. Abbreviations used in Tables 5 and 6 are AF = African Descent, AS = Asian, CA = Caucasian, HL = Hispanic-Latino, and AM = Native American.

Table 5. Frequency of Observed EDG4 Haplotypes In Unrelated Individuals

HAP No.	HAP ID	Total	CA	AF	AS	HL	AM
1	1723271	0.61	0.0	2.5	0.0	0.0	0.0
2	1723269	1.22	0.0	0.0	5.0	0.0	0.0
3	1723270	0.61	0.0	2.5	0.0	0.0	0.0
4	1723265	19.51	23.81	10.0	32.5	11.11	16.67
5	1723264	73.78	76.19	85.0	57.5	80.56	50.0
6	1723266	1.83	0.0	0.0	0.0	2.78	33.33
7	1723272	0.61	0.0	0.0	2.5	0.0	0.0
8	1723267	1.83	0.0	0.0	2.5	5.56	0.0

Table 6. Frequency of Observed EDG4 Haplotype Pairs In Unrelated Individuals

	HAP1	HAP2	Total	CA	AF	AS	HL	AM
5	5	5	54.88	61.9	75.0	25.0	66.67	0.0
	4	4	7.32	9.52	5.0	10.0	5.56	0.0
	5	1	1.22	0.0	5.0	0.0	0.0	0.0
	5	3	1.22	0.0	5.0	0.0	0.0	0.0
	5	7	1.22	0.0	0.0	5.0	0.0	0.0
10	5	4	24.39	28.57	10.0	45.0	11.11	33.33
	5	8	3.66	0.0	0.0	5.0	11.11	0.0
	5	6	3.66	0.0	0.0	0.0	5.56	66.67
	5	2	2.44	0.0	0.0	10.0	0.0	0.0

15       The size and composition of the Index Repository were chosen to represent the genetic diversity across and within four major population groups comprising the general United States population. For example, as described in Table 1 above, this repository contains approximately equal sample sizes of African-descent, Asian-American, European-American, and Hispanic-Latino population groups. Almost all individuals representing each group had all four grandparents with the same

20   ethnogeographic background. The number of unrelated individuals in the Index Repository provides a sample size that is sufficient to detect SNPs and haplotypes that occur in the general population with high statistical certainty. For instance, a haplotype that occurs with a frequency of 5% in the general population has a probability higher than 99.9% of being observed in a sample of 80 individuals from the general population. Similarly, a haplotype that occurs with a frequency of 10% in a specific population

25   group has a 99% probability of being observed in a sample of 20 individuals from that population group. In addition, the size and composition of the Index Repository means that the relative frequencies determined therein for the haplotypes and haplotype pairs of the EDG4 gene are likely to be similar to the relative frequencies of these EDG4 haplotypes and haplotype pairs in the general U.S. population and in the four population groups represented in the Index Repository. The genetic diversity observed

30   for the three Native Americans is presented because it is of scientific interest, but due to the small sample size it lacks statistical significance.

In view of the above, it will be seen that the several advantages of the invention are achieved and other advantageous results attained.

35       As various changes could be made in the above methods and compositions without departing from the scope of the invention, it is intended that all matter contained in the above description and shown in the accompanying drawings shall be interpreted as illustrative and not in a limiting sense.

All references cited in this specification, including patents and patent applications, are hereby incorporated in their entirety by reference. The discussion of references herein is intended merely to summarize the assertions made by their authors and no admission is made that any reference constitutes

40   prior art. Applicants reserve the right to challenge the accuracy and pertinency of the cited references.

## What is Claimed is:

1. A method for haplotyping the endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) gene of an individual, which comprises determining which of the EDG4 haplotypes shown in the table immediately below defines one copy of the individual's EDG4 gene, wherein each of the EDG4 haplotypes comprises a set of polymorphisms whose positions and identities are set forth in the table immediately below:

Haplotype Number <sup>a</sup>								PS	PS
1	2	3	4	5	6	7	8	No. <sup>b</sup>	Pos. <sup>c</sup>
C	T	T	T	T	T	T	T	1	216
C	C	C	C	C	C	T	T	2	395
T	G	G	T	T	T	T	T	3	523
G	A	G	A	G	G	A	G	4	806
A	G	G	G	G	G	G	G	5	812
T	G	G	G	G	G	G	G	6	1291
G	G	G	G	G	G	T	G	7	3099
C	C	C	C	C	T	C	C	8	3449

<sup>a</sup>Alleles for haplotypes are presented 5' to 3' in each column;

<sup>b</sup>PS = polymorphic site;

<sup>c</sup>Position of PS within SEQ ID NO:1.

2. The method of claim 1, wherein the determining step comprises identifying the phased sequence of nucleotides present at each of PS1-PS8 on the one copy of the individual's EDG4 gene.
3. A method for haplotyping the endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) gene of an individual, which comprises determining which of the EDG4 haplotype pairs shown in the table immediately below defines both copies of the individual's EDG4 gene, wherein each of the EDG4 haplotype pairs consists of first and second haplotypes which comprise first and second sets of polymorphisms whose positions and identities are set forth in the table immediately below:

Haplotype Pair <sup>a</sup>								PS	PS
5/5	4/4	5/1	5/3	5/7	5/4	5/8	5/6	No. <sup>b</sup>	Pos. <sup>c</sup>
T/T	T/T	T/C	T/T	T/T	T/T	T/T	T/T	1	216
C/C	C/C	C/C	C/C	C/T	C/C	C/T	C/C	2	395
T/T	T/T	T/T	T/G	T/T	T/T	T/T	T/T	3	523
G/G	A/A	G/G	G/G	G/A	G/A	G/G	G/G	4	806
G/G	G/G	G/A	G/G	G/G	G/G	G/G	G/G	5	812
G/G	G/G	G/T	G/G	G/G	G/G	G/G	G/G	6	1291
G/G	G/G	G/G	G/G	G/T	G/G	G/G	G/G	7	3099
C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T	8	3449

	Haplotype Pair <sup>a</sup>	PS No. <sup>b</sup>	PS Pos. <sup>c</sup>
	5/2		
	T/T	1	216
	C/C	2	395
5	T/G	3	523
	G/A	4	806
	G/G	5	812
	G/G	6	1291
	G/G	7	3099
10	C/C	8	3449

<sup>a</sup>Haplotype pairs are represented as 1<sup>st</sup> haplotype/2<sup>nd</sup> haplotype; with alleles of each haplotype shown 5' to 3' as 1<sup>st</sup> polymorphism/2<sup>nd</sup> polymorphism in each column;

<sup>b</sup>PS = polymorphic site;

<sup>c</sup>Position of PS in SEQ ID NO:1.

4. The method of claim 3, wherein the determining step comprises identifying the phased sequence of nucleotides present at each of PS1-PS8 on both copies of the individual's EDG4 gene.
5. A method for genotyping the endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) gene of an individual, comprising determining for the two copies of the EDG4 gene present in the individual the identity of the nucleotide pair at one or more polymorphic sites (PS) selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7 and PS8, wherein the one or more PS have the location and alternative alleles shown in SEQ ID NO:1.
6. The method of claim 5, wherein the determining step comprises:
  - (a) isolating from the individual a nucleic acid mixture comprising both copies of the EDG4 gene, or a fragment thereof, that are present in the individual;
  - (b) amplifying from the nucleic acid mixture a target region containing the selected polymorphic site;
  - (c) hybridizing a primer extension oligonucleotide to one allele of the amplified target region;
  - (d) performing a nucleic acid template-dependent, primer extension reaction on the hybridized genotyping oligonucleotide in the presence of at least two different terminators of the reaction, wherein said terminators are complementary to the alternative nucleotides present at the selected polymorphic site; and
  - (e) detecting the presence and identity of the terminator in the extended genotyping oligonucleotide.
7. The method of claim 5, which comprises determining for the two copies of the EDG4 gene present in the individual the identity of the nucleotide pair at each of PS1-PS8.
8. A method for haplotyping the endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) gene of an individual which comprises determining, for one copy of the EDG4 gene present in the individual, the identity of the nucleotide at two or more

polymorphic sites (PS) selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7 and PS8, wherein the selected PS have the location and alternative alleles shown in SEQ ID NO:1.

9. The method of claim 8, wherein the determining step comprises:

- (a) isolating from the individual a nucleic acid sample containing only one of the two copies of the EDG4 gene, or a fragment thereof, that is present in the individual;
- (b) amplifying from the nucleic acid sample a target region containing the selected polymorphic site;
- (c) hybridizing a primer extension oligonucleotide to one allele of the amplified target region;
- (d) performing a nucleic acid template-dependent, primer extension reaction on the hybridized genotyping oligonucleotide in the presence of at least two different terminators of the reaction, wherein said terminators are complementary to the alternative nucleotides present at the selected polymorphic site; and
- (e) detecting the presence and identity of the terminator in the extended genotyping oligonucleotide.

10. A method for predicting a haplotype pair for the endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) gene of an individual comprising:

- (a) identifying an EDG4 genotype for the individual, wherein the genotype comprises the nucleotide pair at two or more polymorphic sites (PS) selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7 and PS8, wherein the selected PS have the location and alternative alleles shown in SEQ ID NO:44;
- (b) enumerating all possible haplotype pairs which are consistent with the genotype;
- (c) comparing the possible haplotype pairs to the haplotype pair data set forth in the table immediately below; and
- (d) assigning a haplotype pair to the individual that is consistent with the data

Haplotype Pair <sup>a</sup>								PS	PS
5/5	4/4	5/1	5/3	5/7	5/4	5/8	5/6	No. <sup>b</sup>	Pos. <sup>c</sup>
T/T	T/T	T/C	T/T	T/T	T/T	T/T	T/T	1	216
C/C	C/C	C/C	C/C	C/T	C/C	C/T	C/C	2	395
T/T	T/T	T/T	T/G	T/T	T/T	T/T	T/T	3	523
G/G	A/A	G/G	G/G	G/A	G/A	G/G	G/G	4	806
G/G	G/G	G/A	G/G	G/G	G/G	G/G	G/G	5	812
G/G	G/G	G/T	G/G	G/G	G/G	G/G	G/G	6	1291
G/G	G/G	G/G	G/G	G/T	G/G	G/G	G/G	7	3099
C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T	8	3449

	Haplotype Pair <sup>a</sup>	PS	PS
		No. <sup>b</sup>	Pos. <sup>c</sup>
25	5/2		
	T/T	1	216
	C/C	2	395
	T/G	3	523
	G/A	4	806
	G/G	5	812
30	G/G	6	1291
	G/G	7	3099
	C/C	8	3449

- 35 <sup>a</sup>Haplotype pairs are represented as 1<sup>st</sup> haplotype/2<sup>nd</sup> haplotype; with alleles of each haplotype shown 5' to 3' as 1<sup>st</sup> polymorphism/2<sup>nd</sup> polymorphism in each column;  
<sup>b</sup>PS = polymorphic site;  
<sup>c</sup>Position of PS in SEQ ID NO:1.

11. The method of claim 10, wherein the identified genotype of the individual comprises the nucleotide pair at each of PS1-PS8, which have the location and alternative alleles shown in SEQ ID NO:1.
12. A method for identifying an association between a trait and at least one haplotype or haplotype pair of the endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) gene which comprises comparing the frequency of the haplotype or haplotype pair in a population exhibiting the trait with the frequency of the haplotype or haplotype pair in a reference population, wherein the haplotype is selected from haplotypes 1-8 shown in the table presented immediately below, wherein each of the haplotypes comprises a set of polymorphisms whose positions and identities are set forth in the table immediately below:

	Haplotype Number <sup>a</sup>								PS	PS
	1	2	3	4	5	6	7	8	No. <sup>b</sup>	Pos. <sup>c</sup>
	C	T	T	T	T	T	T	T	1	216
15	C	C	C	C	C	C	T	T	2	395
	T	G	G	T	T	T	T	T	3	523
	G	A	G	A	G	G	A	G	4	806
	A	G	G	G	G	G	G	G	5	812
	T	G	G	G	G	G	G	G	6	1291
20	G	G	G	G	G	G	T	G	7	3099
	C	C	C	C	C	T	C	C	8	3449

- 25 <sup>a</sup>Alleles for haplotypes are presented 5' to 3' in each column;  
<sup>b</sup>PS = polymorphic site;  
<sup>c</sup>Position of PS in SEQ ID NO:1;

- and wherein the haplotype pair is selected from the haplotype pairs shown in the table immediately below, wherein each of the EDG4 haplotype pairs consists of first and second haplotypes which comprise first and second sets of polymorphisms whose positions and identities are set forth in the table immediately below:



	Haplotype Pair <sup>a</sup>								PS	PS
	5/5	4/4	5/1	5/3	5/7	5/4	5/8	5/6	No. <sup>b</sup>	Pos. <sup>c</sup>
35	T/T	T/T	T/C	T/T	T/T	T/T	T/T	T/T	1	216
	C/C	C/C	C/C	C/C	C/T	C/C	C/T	C/C	2	395
	T/T	T/T	T/T	T/G	T/T	T/T	T/T	T/T	3	523
	G/G	A/A	G/G	G/G	G/A	G/A	G/G	G/G	4	806
	G/G	G/G	G/A	G/G	G/G	G/G	G/G	G/G	5	812
40	G/G	G/G	G/T	G/G	G/G	G/G	G/G	G/G	6	1291
	G/G	G/G	G/G	G/G	G/T	G/G	G/G	G/G	7	3099
	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T	8	3449

	Haplotype Pair <sup>a</sup>		PS	PS
	5/2		No. <sup>b</sup>	Pos. <sup>c</sup>
45	T/T		1	216
	C/C		2	395
	T/G		3	523
	G/A		4	806
	G/G		5	812
50	G/G		6	1291
	G/G		7	3099
	C/C		8	3449

<sup>a</sup>Haplotype pairs are represented as 1<sup>st</sup> haplotype/2<sup>nd</sup> haplotype; with alleles of each haplotype shown 5' to 3' as 1<sup>st</sup> polymorphism/2<sup>nd</sup> polymorphism in each column;

<sup>b</sup>PS = polymorphic site;

<sup>c</sup>Position of PS in SEQ ID NO:1;

wherein a higher frequency of the haplotype or haplotype pair in the trait population than in the reference population indicates the trait is associated with the haplotype or haplotype pair.

13. The method of claim 12, wherein the trait is a clinical response to a drug targeting EDG4.
14. An isolated genotyping oligonucleotide for detecting a polymorphism in the endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) gene at a polymorphic site (PS) selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7 and PS8, wherein the selected PS have the location and alternative alleles shown in SEQ ID NO:1.
15. The isolated genotyping oligonucleotide of claim 14, which is an allele-specific oligonucleotide that specifically hybridizes to an allele of the EDG4 gene at a region containing the polymorphic site.
16. The allele-specific oligonucleotide of claim 15, which comprises a nucleotide sequence selected from the group consisting of SEQ ID NOS:4-11, the complements of SEQ ID NOS:4-11, and SEQ ID NOS:12-27.
17. The isolated genotyping oligonucleotide of claim 14, which is a primer-extension oligonucleotide.
18. The primer-extension oligonucleotide of claim 17, which comprises a nucleotide sequence selected from the group consisting of SEQ ID NOS:28-43.

19. A kit for genotyping the endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) gene of an individual, which comprises a set of oligonucleotides designed to genotype each of polymorphic sites (PS) PS1, PS2, PS3, PS4, PS5, PS6, PS7 and PS8, wherein the selected PS have the location and alternative alleles shown in SEQ ID NO:1.
20. An isolated polynucleotide comprising a nucleotide sequence selected from the group consisting of:
- (a) a first nucleotide sequence which comprises a endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) isogene, wherein the EDG4 isogene is selected from the group consisting of isogenes 1- 4 and 6 - 8 shown in the table immediately below and wherein each of the isogenes comprises the regions of SEQ ID NO:1 shown in the table immediately below and wherein each of the isogenes 1- 4 and 6 - 8 is further defined by the corresponding set of polymorphisms whose positions and identities are set forth in the table immediately below

Isogene Number <sup>a</sup>								PS	PS	Region
1	2	3	4	5	6	7	8	No. <sup>b</sup>	Pos. <sup>c</sup>	Examined <sup>d</sup>
C	T	T	T	T	T	T	T	1	216	30-1325
C	C	C	C	C	C	T	T	2	395	30-1325
T	G	G	T	T	T	T	T	3	523	30-1325
G	A	G	A	G	G	A	G	4	806	30-1325
A	G	G	G	G	G	G	G	5	812	30-1325
T	G	G	G	G	G	G	G	6	1291	30-1325
G	G	G	G	G	G	T	G	7	3099	2883-3624
C	C	C	C	C	T	C	C	8	3449	2883-3624

<sup>a</sup>Alleles for isogenes are presented 5' to 3' in each column;

<sup>b</sup>PS = polymorphic site;

<sup>c</sup>Position of PS in SEQ ID NO:1;

<sup>d</sup>Region examined represents the nucleotide positions defining the start and stop positions within SEQ ID NO:1.

- (b) a second nucleotide sequence which comprises a fragment of the first nucleotide sequence, wherein the fragment comprises one or more polymorphisms selected from the group consisting of cytosine at PS1, thymine at PS2, guanine at PS3, adenine at PS4, adenine at PS5, thymine at PS6, thymine at PS7 and thymine at PS8, wherein the selected polymorphism has the location set forth in the table immediately above; and
- (c) a third nucleotide sequence which is complementary to the first or second nucleotide sequence.
21. The isolated polynucleotide of claim 20, which is a DNA molecule and comprises both the first and third nucleotide sequences and further comprises expression regulatory elements operably linked to the first nucleotide sequence.
22. A recombinant nonhuman organism transformed or transfected with the isolated polynucleotide of claim 20, wherein the organism expresses an EDG4 protein encoded by the first nucleotide

sequence.

23. The recombinant nonhuman organism of claim 22, which is a transgenic animal.
24. The isolated polynucleotide of claim 20 which consists of the second nucleotide sequence.
25. An isolated polynucleotide comprising a nucleotide sequence selected from the group consisting of:

- (a) a coding sequence for a endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) isogene wherein the coding sequence is defined by a haplotype selected from the group consisting of 1c-4c and 7c shown in the table immediately below and wherein the coding sequence comprises SEQ ID NO:2 except at each of the polymorphic sites which have the positions and polymorphisms set forth in the table immediately below:

Coding Sequence Haplotype Number <sup>a</sup>					PS	PS
1c	2c	3c	4c	7c	No. <sup>b</sup>	Pos. <sup>c</sup>
T	G	G	T	T	3	125
G	A	G	A	A	4	408
A	G	G	G	G	5	414

<sup>a</sup>Alleles for coding sequence haplotypes are presented 5' to 3' in each column; the numerical portion of the coding sequence haplotype number represents the number of the parent full EDG4 haplotype;

<sup>b</sup>PS = polymorphic site;

<sup>c</sup>Position of PS in SEQ ID NO:2;

and

- (b) a fragment of the coding sequence, wherein the fragment comprises at least one polymorphism selected from the group consisting of guanine at a position corresponding to nucleotide 125, adenine at a position corresponding to nucleotide 408 and adenine at a position corresponding to nucleotide 414, wherein said positions in the coding sequence and the fragment refer to SEQ ID NO:2.
26. A recombinant nonhuman organism transformed or transfected with the isolated polynucleotide of claim 25, wherein the organism expresses a endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) protein encoded by the polymorphic variant sequence.
27. The recombinant nonhuman organism of claim 26, which is a transgenic animal.
28. An isolated polypeptide comprising an amino acid sequence which is a polymorphic variant of a reference sequence for the endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) protein or a fragment thereof, wherein the reference sequence comprises SEQ ID NO:3 and the polymorphic variant comprises one or more variant amino acids selected from the group consisting of glycine at a position corresponding to amino acid position 42.
29. An isolated monoclonal antibody specific for and immunoreactive with the isolated polypeptide of claim 28.

30. A method for screening for drugs targeting the isolated polypeptide of claim 28 which comprises contacting the EDG4 polymorphic variant with a candidate agent and assaying for binding activity.
31. A computer system for storing and analyzing polymorphism data for the endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 gene, comprising:
- (a) a central processing unit (CPU);
  - (b) a communication interface;
  - (c) a display device;
  - (d) an input device; and
  - (e) a database containing the polymorphism data;

wherein the polymorphism data comprises the haplotypes set forth in the table immediately below:

Haplotype Number <sup>a</sup>									PS	PS
1	2	3	4	5	6	7	8		No. <sup>b</sup>	Pos. <sup>c</sup>
C	T	T	T	T	T	T	T		1	216
C	C	C	C	C	C	T	T		2	395
T	G	G	T	T	T	T	T		3	523
G	A	G	A	G	G	A	G		4	806
A	G	G	G	G	G	G	G		5	812
T	G	G	G	G	G	G	G		6	1291
G	G	G	G	G	G	T	G		7	3099
C	C	C	C	C	T	C	C		8	3449

<sup>a</sup>Alleles for haplotypes are presented 5' to 3' in each column;

<sup>b</sup>PS = polymorphic site;

<sup>c</sup>Position of PS in SEQ ID NO:1;

and the haplotype pairs set forth in the table immediately below:

Haplotype Pair <sup>a</sup>									PS	PS
5/5	4/4	5/1	5/3	5/7	5/4	5/8	5/6		No. <sup>b</sup>	Pos. <sup>c</sup>
T/T	T/T	T/C	T/T	T/T	T/T	T/T	T/T		1	216
C/C	C/C	C/C	C/C	C/T	C/C	C/T	C/C		2	395
T/T	T/T	T/T	T/G	T/T	T/T	T/T	T/T		3	523
G/G	A/A	G/G	G/G	G/A	G/A	G/G	G/G		4	806
G/G	G/G	G/A	G/G	G/G	G/G	G/G	G/G		5	812
G/G	G/G	G/T	G/G	G/G	G/G	G/G	G/G		6	1291
G/G	G/G	G/G	G/G	G/T	G/G	G/G	G/G		7	3099
C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T		8	3449

	Haplotype Pair <sup>a</sup>	PS	PS
		No. <sup>b</sup>	Pos. <sup>c</sup>
40	5/2		
	T/T	1	216
	C/C	2	395
	T/G	3	523
	G/A	4	806
45	G/G	5	812
	G/G	6	1291
	G/G	7	3099
	C/C	8	3449

- 50 <sup>a</sup>Haplotype pairs are represented as 1<sup>st</sup> haplotype/2<sup>nd</sup> haplotype; with alleles of each haplotype shown 5' to 3' as 1<sup>st</sup> polymorphism/2<sup>nd</sup> polymorphism in each column;  
<sup>b</sup>PS = polymorphic site;  
<sup>c</sup>Position of PS in SEQ ID NO:1.

32. A genome anthology for the endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (EDG4) gene which comprises EDG4 isogenes defined by any one of haplotypes 1-8 set forth in the table shown below:

	Haplotype Number <sup>a</sup>								PS	PS
	1	2	3	4	5	6	7	8	No. <sup>b</sup>	Pos. <sup>c</sup>
5  10  15	C	T	T	T	T	T	T	T	1	216
	C	C	C	C	C	C	T	T	2	395
	T	G	G	T	T	T	T	T	3	523
	G	A	G	A	G	G	A	G	4	806
	A	G	G	G	G	G	G	G	5	812
	T	G	G	G	G	G	G	G	6	1291
	G	G	G	G	G	G	T	G	7	3099
	C	C	C	C	C	T	C	C	8	3449

- <sup>a</sup>Alleles for haplotypes are presented 5' to 3' in each column;  
<sup>b</sup>PS = polymorphic site;  
<sup>c</sup>Position of PS in SEQ ID NO:1.

1/5

## POLYMORPHISMS IN THE EDG4 GENE

GATCCTCCCT	CATCTCTTCC	CGCAGCTCAC	ACCCCTTTGT	CTTTCCCTAA	
CCCTCCTCTC	CCCACAACCC	AAACCGGGTC	ACACATTGGC	GGCAGGTGGG	100
CTGGGATTCTG	ACCATAGATG	GGTTTTGTTT	CACCTCTGGT	TCAAAAACCT	
ATTGACTTGG	GTGATGTCTT	GCACACAGGT	CTTATTAGTA	GGTAAGAAAC	200
AGCTGCTCCT	CCACCTCCAC	CCCAGCCCCA	TGGAGCTTGC	TTTTTCCAGG	
C					
CCACCGTAGT	CCTCACCACCT	GCCACTTTGC	TTCGAGGTAT	CTGAATTTGA	300
GACCCCTTGT	CTTCCTGCGC	TGACCCTCCT	CTGCAGCTGC	CGCCCCCTCC	
TGTGTCCTTC	CAAGAAGTGC	CACATGCACT	GATCTCTTGT	GTCCCCAGAT	400
T					
[exon 1: 399..					
GGTCATCATG	GGCCAGTGCT	ACTACAACGA	GACCATCGGC	TTCTTCTATA	
ACAACAGTGG	CAAAGAGCTC	AGCTCCCACT	GGCGGCCCAA	GGATGTGGTC	500
GTGGTGGCAC	TGGGGCTGAC	CGTCAGCGTG	CTGGTGCTGC	TGACCAATCT	
G					
GCTGGTCATA	GCAGCCATCG	CCTCCAACCG	CCGCTTCCAC	CAGCCCATCT	600
ACTACCTGCT	CGGCAATCTG	GCCGCGGCTG	ACCTCTTTCG	GGGCGTGGCC	
TACCTCTTCC	TCATGTTCCA	CACCTGGTCCC	CGCACAGCCC	GACTTTCACT	700
TGAGGGCTGG	TTCCTGCGGC	AGGGCTTGCT	GGACACAAGC	CTCACTGCGT	
CGGTGGCCAC	ACTGCTGGCC	ATCGCCGTGG	AGCGGCACCG	CAGTGTGATG	800
GCCGTGCAGC	TGCACAGCCG	CCTGCCCCGT	GGCCGCGTGG	TCATGCTCAT	
A A					
TGTGGGCGTG	TGGGTGGCTG	CCCTGGGCCCT	GGGGCTGCTG	CCTGCCCCACT	900
CCTGGCACTG	CCTCTGTGCC	CTGGACCGCT	GCTCACGCAT	GGCACCCTCTG	
CTCAGCCGCT	CCTATTGGGC	CGTCTGGGCT	CTGTCGAGCC	TGCTTGTCTT	1000
CCTGCTCATG	GTGGCTGTGT	ACACCCGCAT	TTTCTTCTAC	GTGCGGCGGC	
GAGTGCAGCG	CATGGCAGAG	CATGTCAGCT	GCCACCCCCG	CTACCGAGAG	1100
ACCACGCTCA	GCCTGGTCAA	GACTGTTGTC	ATCATCCTGG	GTAAGTGGGG	
..1140]					
CCCCCAGGGG	GCAACAGGGA	CCCTGCTTCC	CCCTATCTCC	CACAAGACAC	1200
TGCGAAGCAA	TGAACAAAC	AAGCTACCTG	ATAGAGGGTG	ATGGGGGATG	
ATTTAGGTGG	GGAATGTCCT	CCAAGGAGGT	GATAATGTAT	GTGAGCAGAG	1300
T					
GCCTGAATGC	TGAGGATAGG	AACAGAGAGC	CTTCCAGGCA	GAGGAAACGG	
AAAGTAAAAA	CCTCAAGGCA	GTACGAGGCC	AGAGCCTCTA	AAGAGGAAAG	1400
TGGGCTTGTG	AGGAGGAGAG	ATCAGCAAGG	CCTTAAGGGG	CACGGACAGG	
AGTCTTATTT	TGAGTTTCAT	GGGAACCCCT	TAAGAGTTTT	TAGCAGTGCT	1500
AGCCGGGCAC	AGTGGCTCAC	GCCTGTAATC	CCAGCACTTT	GGGAGGCCGA	
GGCGGGCAGA	TCAGCTGAGG	TTGGGAGTTC	ACGACCAGCC	TGACCAACAT	1600
GGAGAAACCC	CATCTCTACT	AAAAATGCAA	AATTAGCCAG	GTATGGTGGT	
GTGTGCCCTGT	AATCCCAGCT	ACTCGGGAGC	CTGAGGCAGG	AGAATTGCTT	1700
GAACCCAGGA	GGCGGAGGTT	GTGGTGAGCC	GAGATCACAC	CATTGCACTC	
CAGCCTGGGC	AACAAGAGCG	AAACTCCAGA	ACTCTGTCTC	AGAAAAAATA	1800
AAAAAAAGG	AGTTTTTAGC	AGTGCTAAGC	CTGTTTTAAC	AAAAAGCCTT	
TAAATTGTTT	TTTAATCTAT	ATGTATTTTT	AATTTTTTTT	TTTTTTTTTT	1900
GAGACAGAGT	CTCGCTCTGT	CACCCAGGCT	GGAGTGCAGT	GGCGCAATCT	
CGGCTCACTG	CAACTTCTGC	CTCCAGGTT	CAAGTGGTTC	TCCTGCCTCA	2000
GCCTCCTGAG	TAGCTGGGAC	TACAGGTGCG	TGCCACTATG	CCTGGCTAAT	
TTTTCTATTT	TTAGTAGAGA	CCAGGTTTCA	CTATGTTGGC	CAGGATGGTC	2100
TCGATGTCTT	GACCTCGTGA	TCCACCCGCC	TTGGCCTCCC	AAAGTGCTGG	
GATTACAAGC	ATGAGCCACC	ACGCCCGGCC	CATATTTTTA	AATTTCAAAA	2200
TTTATTTTTG	TTTTTAAGAT	ACAGGATCTT	ACTCTTGCCT	AGGCTGGATT	

FIGURE 1A

2/5

ACAATGGCTC	CATCATAGCT	CACTGCAGCC	TCAACCTTCC	GGGCTCAGGT	2300
GATCCTCCCG	CCTCAGTTTC	CTGAGTAGCT	GGGACTACAG	GCATGTGCCA	
CCATACCTGG	CTAATTTTTT	ATTTTTTGTG	GAAACGAGGT	CTCACTATAT	2400
TGCCTAGGCT	GGTCTTAAAC	TCCTGGCCTC	AAGTGATTGT	CCTGCCTCAG	
CCTGTCAAAT	GCTGCAGTGA	GGTGTGATCG	TGCCGCTTCA	GTGAGGTGTG	2500
ATCGTGCCAC	TGCACCTCCAG	CCTAGGTGAC	AGAAGGAGAC	CCTGTCTCAA	
AAAAAAACAC	ACAGGCCAGG	CGCAGTGGCT	CATGCCTGTA	ATCCCAACAC	2600
TTCCGGGAGG	TGAGGCGGGC	GGATCACGAG	GTCAGGAGAT	CGAGACCATC	
CTGGCCAACA	TGGTGAAACC	CCATCTCTAC	TAAAATACAA	AAAAAAAAAA	2700
ATATAGCCGG	GCATGGTGGC	ACACACCTGT	AGTCCCAGCT	ACTTGGGAGG	
CTGAGGTGGG	GGAATTGCTT	GAACCTGGGA	GGCAGAGGTT	GCAGCAAGCC	2800
AAGATTGTAC	CACTGCACTC	TAGCCTGGCG	ACAGAGCGAG	ACTCCATCTC	
ACACACACAC	ACACACACAC	ACACACACAC	ACACACACAC	ACAAAATGTC	2900
CAGTGCCCTG	GCCACTGTGT	ACCAATTAGC	ATAGATGTCC	CATCTTAGCA	
GGCCTTCCTG	TCACACTGCA	GAATCAGGGA	AACAAGCAGC	TCTTGCCTG	3000
CAAACCTCAC	CACTATTGCT	CGAGCTCTGG	AGACCACTCC	AACCCCAAAC	
TGAGAGCTGT	AGGTGCCGGG	CCTCTGACAG	CTTGCCCTGC	TCACTTCCGC	3100
T					
CTCTGTCCCA	CAGGGGCGTT	CGTGGTCTGC	TGGACACCAG	GCCAGGTGGT	
[exon 2: 3114..					
ACTGCTCCTG	GATGGTTTAG	GCTGTGAGTC	CTGCAATGTC	CTGGCTGTAG	3200
AAAAGTACTT	CCTACTGTTG	GCCGAGGCCA	ACTCACTGGT	CAATGCTGCT	
GTGTACTCTT	GCCGAGATGC	TGAGATGCGC	CGCACCTTCC	GCCGCCTTCT	3300
CTGCTGCGCG	TGCCCTCCGC	AGTCCACCCG	CGAGTCTGTC	CACTATACAT	
CCTCTGCCCA	GGGAGGTGCC	AGCACTCGCA	TCATGCTTCC	CGAGAACGGC	3400
CACCCACTGA	TGGACTCCAC	CCTTTAGCTA	CCTTGAACCT	CAGCGGTACG	
T					
..3427]					
CGGCAAGCAA	CAAATCCACA	GCCCCTGATG	ACTTGTGGGT	GCTCCTGGCT	3500
CAACCCAACC	AACAGGACTG	ACTGACTGGC	AGGACAAGGT	CTGGCATGGC	
ACAGCACCAC	TGCCAGGCCT	CCCCAGGCAC	ACCACTCTGC	CCAGGGAATG	3600
GGGGCTTTGG	GTCATCTCCC	ACTGCCTGGG	GGAGTCAGAT	GGGGTGCAGG	
AATCTGGCTC	TTCAGCCATC	TCAGGTTTAG	GGGGTTTGTA	ACAGACATTA	3700
TTCTGTTTTT	ACTGCGTATC	CTTGGTAAGC	CCTGTGGACT	GGTTCCTGCT	
GTGTGATGCT	GAGGGTTTTA	AGGTGGGGAG	AGATAAGGGC	TCTCTCGGGC	3800
CATGCTACCC	GGTATGACTG	GGTAATGAGG	ACAGACTGTG	GACACCCCAT	
CTACCTGAGT	CTGATTCTTT	AGCAGCAGAG	ACTGAGGGGT	GCAGAGTGTG	3900
AGCTGGGAAA	GGTTTGTGGC	TCCTTGACAG	CTCCAGGGAC	TGGCCTGTCC	
CCGATAGAAT	TGAAGCAGTC	CACGGGGAGG	GGATGATACA	AGGAGTAAAC	4000
CTTTCTTTAC	ACTCTGAGGT	CTCCAAAACA	TTTGTGTGTA	TCAATTCACT	
GTTTCTGTCC	CTGTGTCCCG	GAATTGTCTA	CCTTGGCCCA	AGTCCAGAAA	4100
GGATCTTTCA	ACATGGTATC	TCCTTGGCCC	TGTGCTCGCT	CTTCTTCAGT	
CTGGCTGGCT	GTTATTCCAT	CTCTCTCTTT	TTTTTTTTTT	TTTTTTTTTAG	4200
ACAGAGTCTC	ACTTTGTAGC	CCAGGCTGGA	GAGCAGTAGC	ACAATCTCAG	
CTCGCTGCAA	CTTCTACCTC	CCAGGTTCCT	GATTCTTGTG	CCTCAGGAGA	4300
AGCTGGGATT	ACAGGCGCGT	GCCACCACGT	CCAGCTAATC	TCTGTATTTT	
TATTTTTTAT	ATTTATTTAT	TTGTTTGTGT	ATCTTTGAGA	GGGAGTCTCG	4400
CTCTGTCACC	CAGGCTGGAG	TGCAGTGGCT	TGATCTCAGG	TTACTGCAAC	
TTCTGCCTCC	TGGGTTCAAG	CGATTCTCCT	GCCTCAGCTT	CCCGAATAGC	4500
TGGGTTATAG	GCATCTGCCA	CTACACTCGG	CTAATTTTGT	TATTATAGTA	
GAGATGAGGT	TTCACCAATG	TTGGCCAGGC	TCGTCTCAAA	CTCCTGACCT	4600
TGAATGATCC	ACCCACCTCA	GCTTCCCAAA	GTACTGGGAT	TACAGGTGTG	
AGCTACCGCG	TGCCTGACTA	ATCTTTATAT	TTTGTAGTAG	GATGGGGTCC	4700
CACCATGTTA	GTCAGGCTGG	TCTCGAACTC	TTGACCTCAA	GCGATCCACC	

FIGURE 1B

3/5

TACCCACCTT	GGCCTCCCAA	AGTGCTGGGA	TTATAGGCAT	GAGCCACCAC	4800
GCCTGGCCCC	TCTCTCTTAC	TGGGGTCTTG	AATCTGTTTC	TCTCTGCTTC	
TGTTCTTTGT	TTCCCCCATC	TGTTGCTCTG	CCTCCATCTC	ACTGCCTGTC	4900
TCCCTGTCTT	TCTCTGCAGC	TCAGAGTAGT	GTCTGGACCC	TATGAGCCCC	
CTCCCTCCCC	ACTGCCCTTT	TGTTTGAGCC	CTGATGGTTC	AAGGTCACGT	5000
GCCCAGCCGT	TTCTATCCCA	TTCCCCACCA	CCTGTATTCA	TACTTCCAGC	
CACAGGTTGT	GGGGTTCAGG	TAGTGGGATG	CAGGATTTGA	GCATGACCAT	5100
TTCTCAAGCT	CTTCTGATGT	GCCTGGCATA	CTGGTAAACA	CTTTAGGTGC	
TTTTTTTTTT	AAATTTTTTA	TTTTTTTGTG	ATGGAGTGTC	ACTCTGTCCT	5200
CCAGGCTGGA	GTGTAGTGGT	GCAATCTCGG	CTCGCTGCAA	CCTCCGCCCC	
CGGGTTTCAA	GCGATTCTCC	TGCCTCAGCC	TCCAAGTAG	TTGCGATTAC	5300
AGGCACCTGC	CACCGTGCCC	GGCTAATTTT	TGTATTTTTG	TTAGAGATGG	
GGTTTCACCA	TCTTGGCCAG	GCTTATCTTG	AACTCCCGAC	CTTGTGATCC	5400
ACCCGCCCTG	GCCTCTCAAA	GTGCTGGGAT	TACAGGCATG	AGCCACTGTG	
CTGGGCTTTT	TTTTTTTTTT	TTTTTTTGAGA	CGGAGTCTTG	CTCTGTTTCC	5500
CAGGCTGGAC	TGCAATGACG	TGATCTCAGC	TCACTGCAAC	CTCTGCCTCC	
CAGGTCAAG	TGATTCTCCT	GCATCAGCCT	CCCGGGTAGC	TGGGATTACA	5600
GGGATATGCC	ACCACACCCA	GCTAAATTTT	TTATTTTTAG	TAGAGATGGG	
GTTTCACCAT	TTGGGCCAGG	CTGGTCTCGA	ACTCCTGACC	TCAGGTGAGC	5700
CGCCCGCCTC	AGCCTCTCAA	AGTGCTGGGA	CTATAGGCGT	GAGCCGCCAT	
GCCTGGCCTT	TTTTTTTTTT	TTTTTTTTTT	GAGACAGAGT	CTCACATGAT	5800
CTCAGCTCAC	TGCAACCTCC	GCCCCCTCCA	GGTTCAAGCG	ATTCTTCTGC	
CTCAGCCTCC	CAAGTAGCTG	GGACTACAGG	CATGCACCAC	CATGTCTGGC	5900
TAATTTTTTG	TATTTTAATA	GAGATGGGAT	TTCACCATAT	TGGCCAGGCT	
GGTCTTGAAC	TCCTGACCTC	ATGATCCGCC	TGCTTTGGCC	TCCCAAAGTG	6000
CTGGGATCAG	TCTTCTGAT	TTTTATGTGA	GAACACTGAT	TCAGACAGGT	
TGTGATCTGG	CCCATGGTAA	ACACAGCTTG	GAGATGGCAC	ATACAGGCCC	6100
CGAGGGACCA	GAGAGGCCCC	TAGGAAAGGG	ACTTCAGATT	CCCACGAAGC	
CCAGGGTTGG	TCCCAAGCTC	AGCTGGGCTT	TCCCTACCTC	ACCCACAGC	6200
ACCCAACCCA	TGAGTGCCAG	ATCCATCATG	AGCCCCCTCT	GAGATGCTCA	
GGGAGGCCCC	CTCCAAGGCA	CAGTGTGTGC	CAGGATGTCA	TTCTCTGCTC	6300
CCACCTCACC	AGTGAGGAGG	TGGGGTTATG	TCCCTACTTT	TGAATCCCAT	
GCTGAATAAC	AGCTAGAATT	TGAACCCAGG	TCACTCAGAG	ATGTGAGCCG	6400
CCATGCCCAG	CCACCTGTGC	CAAATTCTTA	AACCACTGAA	ATATGCATTG	
TCAAAGAACC	CAAACTCCC	AGGTCCTTCT	ATATTAAATG	GCCCTGGCCG	6500
AGCACTTTGA	GTGCCAAGGC	TGCCAACTAA	TTTCTATGCA	GTAGAGACCC	
CTGATGACCT	CCAAATGGCA	ACAGTGGGAC	GGGTGAGGTG	GCTCATACCT	6600
GTAATCCCAG	CACTTTGGGC	GGCCGAGGCA	GGCAGATCAC	CTGAGGTCAG	
AAGTTTGAAA	CTAGCCTGGC	CAACATGGTG	AAACCTCATT	TCTACTAAAA	6700
ATGCAAAAAT	TAGCCGGGCG	TGGTGGTGGG	TGCCTGTAAG	CTACTCAGGA	
GGAGGCTAAG	GCAGGAGAAG	TGCTTGAACC	CAGGAGGTGG	AGGTTGCAGT	6800
GAGCCAAGGT	CACGCCACTA	CACTCCAGCC	TGGGCAACAG	AGTAAGATTT	
CTGTCTAAAA	AAAAAAAAAA	AAAAAAAAAGG	CAACTCTGTC	TCAGGCGCTC	6900
TGAGCTGGCT	GTTACATCA	GCTCCTCCAG	GTTTTAACCT	GCACCCCGTG	
GGTCTCCTTC	GAGACCATTT	CCTCTTGTCG	ATCTCCCGTC	TCCACCAAGT	7000
GGCTTCAAAT	TTTTGTCTCT	GCTGGTCCCT	GAAGCTTTCA	GTGCCCACCC	
CTCCAAGTTC	TCTGAAGGAT	ACTACTTTGT	TGGGGTTGGG	GGCAGGTTTT	7100
GCTCTCTCAC	CCAGGCTGGA	GTGCAGTGGC	ACAAACACAG	CTCACTGTAG	
CCTCGACCTC	CTGAGCCCAA	CTTATCCTCC	CGCCTCAGCC	TCCCAAGTAG	7200
CTGGGACCAT	GACCAGCTAA	TTTAGCTATT	TTTTTAATTT	TTATTTTTGT	
AGAGATGGGG	TCTCACTATG	TTCCCCAGAC	TGGTCTTGAA	CTCCTGGGCT	7300
CAAGTAATCT	CCCACCTCAG	ACTCCCAAAG	TGTTGGGATC	ATAGGTGTCA	
GCCACTGCAC	CCAGCCTGAA	GGCCACTACT	GTTCAGGCCT	CCTGGGCAGT	7400
GCTGTGCTTG	TGGGTGACCT	TTTGAA			7426

FIGURE 1C



4/5

## POLYMORPHISMS IN THE CODING SEQUENCE OF EDG4

ATGGTCATCA	TGGGCCAGTG	CTACTACAAC	GAGACCATCG	GCTTCTTCTA	
TAACAACAGT	GGCAAAGAGC	TCAGCTCCCA	CTGGCGGCCC	AAGGATGTGG	100
TCGTGGTGGC	ACTGGGGCTG	ACCGTCAGCG	TGCTGGTGCT	GCTGACCAAT	
		G			
CTGCTGGTCA	TAGCAGCCAT	CGCCTCCAAC	CGCCGCTTCC	ACCAGCCCAT	200
CTACTACCTG	CTCGGCAATC	TGGCCGCGGC	TGACCTCTTC	GCGGGCGTGG	
CCTACCTCTT	CCTCATGTTC	CACACTGGTC	CCCGCACAGC	CCGACTTTCA	300
CTTGAGGGCT	GGTTCCTGCG	GCAGGGCTTG	CTGGACACAA	GCCTCACTGC	
GTCGGTGGCC	AACTGCTGG	CCATCGCCGT	GGAGCGGCAC	CGCAGTGTGA	400
TGGCCGTGCA	GCTGCACAGC	CGCCTGCCCC	GTGGCCGCGT	GGTCATGCTC	
A	A				
ATTGTGGGCG	TGTGGGTGGC	TGCCCTGGGC	CTGGGGCTGC	TGCCTGCCCA	500
CTCCTGGCAC	TGCCTCTGTG	CCCTGGACCG	CTGCTCACGC	ATGGCACCCC	
TGCTCAGCCG	CTCCTATTTG	GCCGTCGGG	CTCTGTCGAG	CCTGCTTGTC	600
TTCCTGCTCA	TGGTGGCTGT	GTACACCCGC	ATTTCTTCT	ACGTGCGGCG	
GCGAGTGCAG	CGCATGGCAG	AGCATGTCAG	CTGCCACCCC	CGCTACCGAG	700
AGACACGCT	CAGCCTGGTC	AAGACTGTTG	TCATCATCCT	GGGGGCGTTC	
GTGGTCTGCT	GGACACCAGG	CCAGGTGGTA	CTGCTCCTGG	ATGGTTTAGG	800
CTGTGAGTCC	TGCAATGTCC	TGGCTGTAGA	AAAGTACTTC	CTACTGTTGG	
CCGAGGCCAA	CTCACTGGTC	AATGCTGCTG	TGTACTCTTG	CCGAGATGCT	900
GAGATGCGCC	GCACCTTCCG	CCGCCTTCTC	TGCTGCGCGT	GCCTCCGCCA	
GTCCACCCGC	GAGTCTGTCC	ACTATACATC	CTCTGCCCAG	GGAGGTGCCA	1000
GCACTCGCAT	CATGCTTCCC	GAGAACGGCC	ACCCACTGAT	GGACTCCACC	
CTTTAG					1056

FIGURE 2

5/5

## ISOFORMS OF THE EDG4 PROTEIN

MVIMGQCYYN	ETIGFFYNNS	GKELSSHWRP	KDVVVVALGL	TVSVLVLLTN	
			G		
LLVIAAIASN	RRFHQPIYYL	LGNLAAADLF	AGVAYLFLMF	HTGPRTARLS	100
LEGWFLRQGL	LDTSLTASVA	TLLAIAVERH	RSVMAVQLHS	RLPRGRVVML	
IVGVVVAALG	LGLLPAHSWH	CLCALDRCR	MAPLLSRSYL	AVWALSSLLV	200
FLLMVAVYTR	IFFYVRRRVQ	RMAEHVSCHP	RYRETTLSLV	KTVVIILGAF	
VVCWTPGQVV	LLLDGLGCES	CNVLAVEKYF	LLAEANSLV	NAAVYSCRDA	300
EMRRTFRRL	CCACLRQSTR	ESVHYTSSAQ	GGASTRIMLP	ENGHPLMDST	
L					351

FIGURE 3

## SEQUENCE LISTING

<110> Genaissance Pharmaceuticals, Inc.  
Kazemi, Amir  
Koshy, Beena  
Sanchis, Angela

<120> HAPLOTYPES OF THE EDG4 GENE

<130> MWH-1001PCT EDG4

<140> tba

<141> 2001-08-06

<150> 60/223,177

<151> 2000-08-04

<160> 44

<170> PatentIn Ver. 2.1

<210> 1

<211> 7426

<212> DNA

<213> Homo sapiens

<220>

<221> allele

<222> (216)

<223> PS1: polymorphic base T or C

<220>

<221> allele

<222> (395)

<223> PS2: polymorphic base C or T

<220>

<221> allele

<222> (523)

<223> PS3: polymorphic base T or G

<220>

<221> allele

<222> (806)

<223> PS4: polymorphic base G or A

<220>

<221> allele

&lt;222&gt; (812)

&lt;223&gt; PS5: polymorphic base G or A

&lt;220&gt;

&lt;221&gt; allele

&lt;222&gt; (1291)

&lt;223&gt; PS6: polymorphic base G or T

&lt;220&gt;

&lt;221&gt; allele

&lt;222&gt; (3099)

&lt;223&gt; PS7: polymorphic base G or T

&lt;220&gt;

&lt;221&gt; allele

&lt;222&gt; (3449)

&lt;223&gt; PS8: polymorphic base C or T

&lt;400&gt; 1

gatcctccct catctcttcc cgcagctcac acccctttgt ctttccctaa ccctcctctc 60  
 ccacaaacc aaacggggtc acacattggc ggcagggtgg ctgggattcg accatagatg 120  
 ggttttgttt cacttctggt tcaaaaacct attgacttgg gtgatgtctt gcacacaggt 180  
 cttattagta ggtaagaaac agctgctcct ccaccyccac ccagcccca tggagcttgc 240  
 tttttccagg ccaccgtagt cctcaccact gccactttgc ttcgagggtat ctgaatttga 300  
 gacccttgc cttcctgcgc tgaccctcct ctgcagctgc cggccctccc tgtgtccttc 360  
 caagaagtgc cacatgcact gatctcttgt gtccycagat ggatcatcat ggccagtgtc 420  
 actacaacga gaccatcggc ttcttctata acaacagtgg caaagagctc agctcccact 480  
 ggcgcccaa ggatgtggtc gtggtggcac tggggctgac cgkcagcgtg ctggtgctgc 540  
 tgaccaatct gctggtcata gcagccatcg cctccaaccg ccgcttcac cagcccatct 600  
 actacctgct cggcaatctg gccgcggtg acctcttcgc gggcggtggc tacctcttcc 660  
 tcatgttcca cactggtccc cgcacagccc gactttcact tgagggtgg ttctgcggc 720  
 agggcttgc ggacacaagc ctactgcgt cgggtggccac actgctggcc atcgccgtgg 780  
 agcgccaccg cagtgtgatg gccgtcagc tccacagccg cctgccccgt ggccgctgg 840  
 tcatgctcat tgtggcggtg tgggtggctg ccctgggcct ggggctgctg cctgcccact 900  
 cctggcactg cctctgtgcc ctggaccgt gtcacgcat ggcacccctg ctcagccgct 960  
 cctatttggc cgtctgggct ctgtcgagcc tgcttctctt cctgctcatg gtggctgtgt 1020  
 acaccgcat tttcttctac gtgcggcgcc gactgcagcg catggcagag catgtcagct 1080  
 gccaccccg ctaccgagag accacgctca gcctggtcaa gactgttgc atcatcctgg 1140  
 gtaagtggg cccccaggg gcaacaggga ccctgcttcc ccctatctcc cacaagacac 1200  
 tgcaagcaa tgaacaaaac aagctacctg atagagggtg atgggggatg atttaggtgg 1260  
 ggaatgtcct ccaaggaggt gataatgtat ktgagcagag gcctgaatgc tgaggatagg 1320  
 aacagagagc cttccaggca gaggaacgg aaagtaaaa cctcaaggca gtacgaggcc 1380  
 agagcctcta aagaggaaag tgggcttgtg aggaggagag atcagcaagg ccttaagggg 1440  
 cacggacagg agtcttattt tgagtttcat gggaaccct taagagtttt tagcagtgtc 1500  
 agccgggac agtggctcac gcctgtaatc ccagcacttt gggaggccga ggccggcaga 1560  
 tcagctgagg ttgggagttc acgaccagcc tgaccaacat ggagaaacc catctctact 1620  
 aaaaatgcaa aattagccag gtatggtgg gtgtgcctgt aatcccagct actcgggagc 1680  
 ctgaggcagg agaattgctt gaaccaggga ggcggaggtt gtggtgagcc gagatcacac 1740

cattgcactc cagcctgggc aacaagagcg aaactccaga actctgtctc agaaaaaaaa 1800  
aaaaaaaaagg agtttttagc agtgctaagc ctgttttaac aaaaagcctt taaattgttt 1860  
tttaatctat atgtattttt aatttttttt tttttttttt gagacagagt ctcgctctgt 1920  
caccagget ggagtgcagt ggcgcaatct cggctcactg caacttctgc ctcccaggtt 1980  
caagtgggtc tcctgcctca gcctcctgag tagctgggac tacaggtgcg tgccactatg 2040  
cctggctaatt ttttctatit ttagtagaga ccaggtttca ctatgttggc caggatgggtc 2100  
tcgatgtctt gacctcgtga tccaccgccc ttggcctccc aaagtgtctg gattacaagc 2160  
atgagccacc acgcccggcc catattttta aatttcaaaa tttatttttg tttttaagat 2220  
acaggatctt actcttgccct aggtctggatt acaatggctc catcatagct cactgcagcc 2280  
tcaaccttcc gggctcaggt gatcctccc cctcagtttc ctgagtagct gggactacag 2340  
gcatgtgcca ccatacctgg ctaatttttt attttttgta gaaacgaggt ctactatat 2400  
tgctaggtct ggtcttaaac tcctggcctc aagtgtattgt cctgcctcag cctgtcaaat 2460  
gctgcagtga ggtgtgatcg tgccgttcca gtgaggtgtg atcgtgccac tgcactccag 2520  
cctaggtgac agaaggagac cctgtctcaa aaaaaaacac acaggccagg cgagtggtct 2580  
catgcctgta atcccaacac ttcgggagggc tgaggcgggc ggatcacgag gtcaggagat 2640  
cgagaccatc ctggccaaca tggtgaaacc ccactctctac taaaatacaa aaaaaaaaaa 2700  
atatagccgg gcatgtggc acacacctgt agtcccagct acttgggagg ctgaggtggg 2760  
ggaattgctt gaacctggga ggcagaggtt gcagcaagcc aagattgtac cactgcactc 2820  
tagcctggcg acagagcgag actccatctc acacacacac acacacacac acacacacac 2880  
acacacacac acaaaatgtc cagtgccttg gccactgtgt accaattagc atagatgtcc 2940  
catcttagca ggccttctcg tcacactgca gaatcaggga aacaagcagc tcttgactg 3000  
caaacttcac cactattgct cgagctctgg agaccactcc aaccccaaac tgagagctgt 3060  
aggtgccggg cctctgacag cttgccctgc tcaactcckc ctctgtccca caggggcgtt 3120  
cgtggtctgc tggacaccag gccaggtggt actgctcctg gatggtttag gctgtgagtc 3180  
ctgcaatgtc ctggctgtag aaaagtaact cctactgttg gccgaggcca actcactggt 3240  
caatgctgct gtgtactctt gccgagatgc tgagatggc cgcaccttcc gccgccttct 3300  
ctgctgcgcg tgccctccgc agtccaccgc cgagtctgtc cactatacat cctctgcca 3360  
gggaggtgcc agcactcgca tcatgcttcc cgagaacggc caccactga tggactccac 3420  
ccttttagcta ccttgaactt cagcgttayg cggcaagcaa caaatccaca gccctgatg 3480  
acttgtgggt gctcctggct caacccaacc aacaggactg actgactggc aggacaaggt 3540  
ctggcatggc acagcaccac tgccaggcct cccaggcac accactctgc ccagggaatg 3600  
ggggcttttg gtcactctcc actgcctggg ggagtcagat ggggtgcagg aatctggctc 3660  
ttcagccatc tcaggtttag ggggtttgta acagacatta ttctgttttc actgcgtatc 3720  
cttggttaagc cctgtggact ggttctctgt gtgtgatgct gagggtttta aggtggggag 3780  
agataagggc tctctcgggc catgtacccc ggtatgactg ggtaatgagg acagactgtg 3840  
gacaccccat ctacctgagt ctgattcttt agcagcagag actgaggggt gcagagtgtg 3900  
agctgggaaa ggtttgtggc tccttgcaac ctccagggac tggcctgtcc ccgatagaat 3960  
tgaagcagtc cacggggagg ggatgataca aggagtaaac ctttctttac actctgaggt 4020  
ctccaaaaca tttgttgta tcaattcagt gtttctgtcc ctgtgtcccg gaattgtcta 4080  
ccttgcccca agtcagaaa ggatctttca acatggtatc tccttgcccc tgtgctcgct 4140  
cttcttcagt ctggctggct gttattccat ctctctcttt tttttttttt ttttttttag 4200  
acagagtctc actttgtagc ccaggctgga gagcagtagc acaatctcag ctgctgcaa 4260  
cttctacctc ccaggttcca gattcttggt cctcaggaga agctgggatt acaggcgcgt 4320  
gccaccacgt ccagctaate tctgtatttt tatttttatt atttatttat ttgtttgttt 4380  
atctttgaga gggagtctcg ctctgtcacc caggctggag tgcaagtggct tgatctcagg 4440  
ttactgcaac ttctgcctcc tgggttcaag cgattctcct gcctcagctt cccgaatagc 4500  
tgggttatag gcatctgcca ctacactcgg ctaatttttg tattatagta gagatgaggt 4560  
ttcaccaatg ttggccaggc tcgtctcaaa ctctgacct tgaatgatcc acccaacctca 4620

```

gcttcccaaa gtactgggat tacaggtgtg agctaccgcg tgcctgacta atctttatat 4680
ttttagtaga gatgggggtcc caccatgtta gtcaggctgg tctcgaactc ttgacctcaa 4740
gcgatccacc taccacactt ggccctccaa agtgctggga ttataggcat gagccaccac 4800
gcctggcccc tctctcttac tggggtcttg aatctgtttc tctctgcttc tgttctttgt 4860
ttccccatc tgttgcctcg cctccatctc actgcctgtc tccctgtctt tctctgcagc 4920
tcagagtagt gtctggaccc tatgagcccc ctccctcccc actgcccttt tgtttgagcc 4980
ctgatgggtc aaggtcacgt gccagccgt ttctatccca ttccccacca cctgtattca 5040
tacttccagc cacaggttgt ggggttcagg tagtgggatg caggatttga gcatgacct 5100
ttctcaagct ctctgatgt gcctggcata ctggtaaaca ctttaggtgc tttttttttt 5160
aaatttttta tttttttgtg atggagtgtc actctgtcct ccaggctgga gtgtagtgg 5220
gcaatctcgg ctcgctgcaa cctccgcccc cggttttcaa gcgattctcc tgcctcagcc 5280
tcccaagtag ttgcgattac aggcacctgc caccgtgccc ggctaatttt tgtatttttg 5340
ttagagatgg ggttccacca tcttggccag gcttatcttg aactcccgac cttgtgatcc 5400
accgcctcg gcctctcaaa gtgctgggat tacaggcatg agccactgtg ctgggctttt 5460
tttttttttt ttttttgaga cggagtcttg ctctgtttcc caggctggac tgcaatgacg 5520
tgatctcagc tcactgcaac ctctgcctcc caggttcaag tgattctcct gcatcagcct 5580
ccgggttagc tgggattaca gggtatgcc accacacca gctaaatttt ttatttttag 5640
tagagatggg gtttccacct gtggccagg ctggtctcga actcctgacc tcaggtgagc 5700
cgcccgctc agcctctcaa agtctggga ctataggcgt gagccgcat gcctggcctt 5760
tttttttttt tttttttttt gagacagagt ctccatgat ctgagctcac tgcaacctcc 5820
gccccctcca ggttcaagcg attcttctgc ctgagctcc caagtagctg ggactacagg 5880
catgcaccac catgtctggc taattttttg tattttaata gagatgggat ttcaccatat 5940
tggccaggct ggtcttgaac tctgacctc atgatccgcc tgctttggcc tcccaaagt 6000
ctgggatcag tcttctgat ttttatgtga gaacctgat tcagacaggt tgtgatctgg 6060
cccatggtaa acacagcttg gagatggcac atacaggccc cgagggacca gagaggcccc 6120
taggaaaggg acttcagatt cccacgaagc ccagggttg tcccaagctc agctgggctt 6180
tccctacctc accccacagc acccaacca tgagtgccag atccatcatg agccccctct 6240
gagatgtca gggaggcccc ctccaaggca cagtgtgtgc caggatgtca ttctctgctc 6300
ccacctcacc agtgaggagg tggggttatg tccctacttt tgaatcccat gctgaataac 6360
agctagaatt tgaaccagg tcactcagag atgtgagccg ccatgccag ccacctgtgc 6420
caaattctta aacctgaa atatgcattg tcaaagaacc caaaactccc aggtccttct 6480
atattaaatg gccctggccg agcactttga gtgccaaggc tgccaactaa tttctatgca 6540
gtagagacc ctgatgacct ccaaattgga acagtgggac gggtaggtg gctcatacct 6600
gtaatccag cactttgggc ggccgaggca ggcagatcac ctgaggtcag aagtttgaaa 6660
ctagcctggc caacatggtg aaacctcatt tctactaaaa atgcaaaaat tagccgggcg 6720
tgggtgggg tgcctgtaag ctactcagga ggaggctaag gcaggagaag tgcttgaacc 6780
caggaggtgg aggttgagc gagccaaggt cacgccacta cactccagcc tgggcaacag 6840
agtaagattt ctgtctaaaa aaaaaaaaaa aaaaaaagg caactctgtc tcaggcgctc 6900
tgagctggct gttcacatca gctcctccag gttttaacct gcaccccggt ggtctccttc 6960
gagaccaatt cctctgtcg atctccgctc tccaccaagt ggcttcaaat tttgtctct 7020
gctggtccct gaagcttca gtgcccaccc ctccaagttc tctgaaggat actactttgt 7080
tggggtggg ggcaggtttt gctctctcac ccaggctgga gtgcagtggc acaaacacag 7140
ctcactgtag cctcgacctc ctgagcccaa cttatcctcc cgcctcagcc tcccaagtag 7200
ctgggacct gaccagctaa tttagctatt tttttaattt ttatttttgt agagatgggg 7260
tctcactatg ttccccagac tggcttgaa ctctgggct caagtaatct cccacctcag 7320
actcccaaag tgttgggatc ataggtgtca gccactgcac ccagcctgaa ggccactact 7380
gttcaggcct cctgggcagt gctgtgcttg tgggtgacct tttgaa 7426

```

<210> 2  
 <211> 1056  
 <212> DNA  
 <213> Homo sapiens

<400> 2  
 atgggtcatca tgggccagtg ctactacaac gagaccatcg gcttcttcta taacaacagt 60  
 ggcaaagagc tcagctcccā ctggcgggccc aaggatgtgg tcgtggtggc actggggctg 120  
 accgtcagcg tgctggtgct gctgaccaat ctgctggtca tagcagccat cgcctccaac 180  
 cgccgcttcc accagcccat ctactacctg ctcggaatc tggccgaggc tgacctcttc 240  
 gcgggcgtgg cctacctctt cctcatgttc cacactggtc cccgcacagc ccgactttca 300  
 cttgagggtt ggttcctgcg gcagggtctt ctggacacaa gcctcactgc gtcggtggcc 360  
 acactgctgg ccatcgccgt ggagcggcac cgcagtgtga tggccgtgca gctgcacagc 420  
 cgcctgcccc gtggccgctt ggatcatgtc attgtgggct tggggtggc tgcctgggc 480  
 ctggggtgct tgctgcca ctctggcac tgcctctgtg ccctggaccg ctgctcacgc 540  
 atggcaccct tgctcagccg ctctattttg gccgtctggg ctctgtcgag cctgcttgtc 600  
 ttctgtctca tgggtgctgt gtacaccgcg atttcttct acgtgcggcg gcgagtgcag 660  
 cgcattggcag agcatgtcag ctgccacccc cgtaccgag agaccacgct cagcctggtc 720  
 aagactgttg tcatcatcct gggggcgctt gtggtctgct ggacaccagg ccagggtgga 780  
 ctgctcctgg atggttttag ctgtgagtc tgcaatgtcc tggctgtaga aaagtacttc 840  
 ctactgttgg ccgaggccaa ctactggtc aatgctgctg tgtactcttg ccgagatgct 900  
 gagatgcgcc gcaccttcg ccgcttctc tgcgtgcgct gcctccgcca gtccaccgc 960  
 gagtctgtcc actatacatc ctctgccag ggaggtgcca gcactcgcat catgcttccc 1020  
 gagaacggcc acccactgat ggactccacc ctttag 1056

<210> 3  
 <211> 351  
 <212> PRT  
 <213> Homo sapiens

<400> 3  
 Met Val Ile Met Gly Gln Cys Tyr Tyr Asn Glu Thr Ile Gly Phe Phe  
 1 5 10 15  
 Tyr Asn Asn Ser Gly Lys Glu Leu Ser Ser His Trp Arg Pro Lys Asp  
 20 25 30  
 Val Val Val Val Ala Leu Gly Leu Thr Val Ser Val Leu Val Leu Leu  
 35 40 45  
 Thr Asn Leu Leu Val Ile Ala Ala Ile Ala Ser Asn Arg Arg Phe His  
 50 55 60  
 Gln Pro Ile Tyr Tyr Leu Leu Gly Asn Leu Ala Ala Asp Leu Phe  
 65 70 75 80

Ala Gly Val Ala Tyr Leu Phe Leu Met Phe His Thr Gly Pro Arg Thr  
85 90 95

Ala Arg Leu Ser Leu Glu Gly Trp Phe Leu Arg Gln Gly Leu Leu Asp  
100 105 110

Thr Ser Leu Thr Ala Ser Val Ala Thr Leu Leu Ala Ile Ala Val Glu  
115 120 125

Arg His Arg Ser Val Met Ala Val Gln Leu His Ser Arg Leu Pro Arg  
130 135 140

Gly Arg Val Val Met Leu Ile Val Gly Val Trp Val Ala Ala Leu Gly  
145 150 155 160

Leu Gly Leu Leu Pro Ala His Ser Trp His Cys Leu Cys Ala Leu Asp  
165 170 175

Arg Cys Ser Arg Met Ala Pro Leu Leu Ser Arg Ser Tyr Leu Ala Val  
180 185 190

Trp Ala Leu Ser Ser Leu Leu Val Phe Leu Leu Met Val Ala Val Tyr  
195 200 205

Thr Arg Ile Phe Phe Tyr Val Arg Arg Arg Val Gln Arg Met Ala Glu  
210 215 220

His Val Ser Cys His Pro Arg Tyr Arg Glu Thr Thr Leu Ser Leu Val  
225 230 235 240

Lys Thr Val Val Ile Ile Leu Gly Ala Phe Val Val Cys Trp Thr Pro  
245 250 255

Gly Gln Val Val Leu Leu Leu Asp Gly Leu Gly Cys Glu Ser Cys Asn  
260 265 270

Val Leu Ala Val Glu Lys Tyr Phe Leu Leu Leu Ala Glu Ala Asn Ser  
275 280 285

Leu Val Asn Ala Ala Val Tyr Ser Cys Arg Asp Ala Glu Met Arg Arg  
290 295 300

Thr Phe Arg Arg Leu Leu Cys Cys Ala Cys Leu Arg Gln Ser Thr Arg  
305 310 315 320

Glu Ser Val His Tyr Thr Ser Ser Ala Gln Gly Gly Ala Ser Thr Arg  
325 330 335



Ile Met Leu Pro Glu Asn Gly His Pro Leu Met Asp Ser Thr Leu  
340 345 350

<210> 4

<211> 15

<212> DNA

<213> Homo sapiens

<400> 4

ctccaccycc acccc

15

<210> 5

<211> 15

<212> DNA

<213> Homo sapiens

<400> 5

tgtgtccyca gatgg

15

<210> 6

<211> 15

<212> DNA

<213> Homo sapiens

<400> 6

ctgaccgkca gcgtg

15

<210> 7

<211> 15

<212> DNA

<213> Homo sapiens

<400> 7

tggccgtrca gctgc

15

<210> 8

<211> 15

<212> DNA

<213> Homo sapiens

<400> 8

tgcagctrca cagcc

15

<210> 9  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 9  
aatgtatktg agcag

15

<210> 10  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 10  
cacttcckcc tctgt

15

<210> 11  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 11  
agcggtaygc ggcaa

15

<210> 12  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 12  
ctgctcctcc accyc

15

<210> 13  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 13  
ggggctgggg tggrg

15

<210> 14  
<211> 15

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;400&gt; 14

atctcttgtg tccyc

15

&lt;210&gt; 15

&lt;211&gt; 15

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;400&gt; 15

tgatgaccat ctgrg

15

&lt;210&gt; 16

&lt;211&gt; 15

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;400&gt; 16

ctggggctga ccgkc

15

&lt;210&gt; 17

&lt;211&gt; 15

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;400&gt; 17

caccagcacg ctgmc

15

&lt;210&gt; 18

&lt;211&gt; 15

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;400&gt; 18

gtgtgatggc cgttc

15

&lt;210&gt; 19

&lt;211&gt; 15

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

<400> 19  
ggctgtgcag ctgya

15

<210> 20  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 20  
tggccgtgca gctrc

15

<210> 21  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 21  
gcaggcggct gtgya

15

<210> 22  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 22  
ggtgataatg tatkt

15

<210> 23  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 23  
aggcctctgc tcama

15

<210> 24  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 24  
cctgctcact tccckc

15

<210> 25  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 25  
tgtgggacag aggmg

15

<210> 26  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 26  
aacttcagcg gtayg

15

<210> 27  
<211> 15  
<212> DNA  
<213> Homo sapiens

<400> 27  
tgttgcttgc cgcrt

15

<210> 28  
<211> 10  
<212> DNA  
<213> Homo sapiens

<400> 28  
ctcctccacc

10

<210> 29  
<211> 10  
<212> DNA  
<213> Homo sapiens

<400> 29  
gctggggtgg

10

<210> 30  
<211> 10

<212> DNA

<213> Homo sapiens

<400> 30

tcttgtgtcc

10

<210> 31

<211> 10

<212> DNA

<213> Homo sapiens

<400> 31

tgaccatctg

10

<210> 32

<211> 10

<212> DNA

<213> Homo sapiens

<400> 32

gggctgaccg

10

<210> 33

<211> 10

<212> DNA

<213> Homo sapiens

<400> 33

cagcacgctg

10

<210> 34

<211> 10

<212> DNA

<213> Homo sapiens

<400> 34

tgatggccgt

10

<210> 35

<211> 10

<212> DNA

<213> Homo sapiens

<400> 35  
tgtgcagctg

10

<210> 36  
<211> 10  
<212> DNA  
<213> Homo sapiens

<400> 36  
ccgtgcagct

10

<210> 37  
<211> 10  
<212> DNA  
<213> Homo sapiens

<400> 37  
ggcggctgtg

10

<210> 38  
<211> 10  
<212> DNA  
<213> Homo sapiens

<400> 38  
gataatgtat

10

<210> 39  
<211> 10  
<212> DNA  
<213> Homo sapiens

<400> 39  
cctctgctca

10

<210> 40  
<211> 10  
<212> DNA  
<213> Homo sapiens

<400> 40  
gtcacttcc

10

<210> 41  
<211> 10  
<212> DNA  
<213> Homo sapiens

<400> 41  
gggacagagg

10

<210> 42  
<211> 10  
<212> DNA  
<213> Homo sapiens

<400> 42  
ttcagcggta

10

<210> 43  
<211> 10  
<212> DNA  
<213> Homo sapiens

<400> 43  
tgcttgccgc

10

<210> 44  
<211> 960  
<212> DNA  
<213> Homo sapiens

<220>  
<221> allele  
<222> (30)  
<223> PS1: polymorphic base T or C

<220>  
<221> misc\_feature  
<222> (61)..(120)  
<223> N's represent sequence between PS1 and PS2

<220>  
<221> allele  
<222> (150)  
<223> PS2: polymorphic base C or T



<220>  
<221> misc\_feature  
<222> (181)..(240)  
<223> N's represent sequence between PS2 and PS3

<220>  
<221> allele  
<222> (270)  
<223> PS3: polymorphic base T or G

<220>  
<221> misc\_feature  
<222> (301)..(360)  
<223> N's represent sequence between PS3 and PS4

<220>  
<221> allele  
<222> (390)  
<223> PS4: polymorphic base G or A

<220>  
<221> misc\_feature  
<222> (421)..(480)  
<223> N's represent sequence between PS4 and PS5

<220>  
<221> allele  
<222> (510)  
<223> PS5: polymorphic base G or A

<220>  
<221> misc\_feature  
<222> (541)..(600)  
<223> N's represent sequence between PS5 and PS6

<220>  
<221> allele  
<222> (630)  
<223> PS6: polymorphic base G or T

<220>  
<221> misc\_feature  
<222> (661)..(720)  
<223> N's represent sequence between PS6 and PS7

<220>  
<221> allele  
<222> (750)

<223> PS7: polymorphic base G or T

<220>

<221> misc\_feature

<222> (781)..(840)

<223> N's represent sequence between PS7 and PS8

<220>

<221> allele

<222> (870)

<223> PS8: polymorphic base C or T

<220>

<221> misc\_feature

<222> (901)..(960)

<223> N's represent sequence following PS8

<400> 44

```

agtaggtaag aaacagctgc tcctccaccy ccacccagc cccatggagc ttgctttttc 60
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 120
agtgccacat gcactgatct cttgtgtccy cagatgggtca tcatgggcca gtgctactac 180
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 240
tgtggtcgtg gtggcactgg ggctgaccgk cagcgtgctg gtgctgctga ccaatctgct 300
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 360
gtggagcggc accgcagtgt gatggcctgtr cagctgcaca gccgcctgcc ccgtggccgc 420
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 480
cggcaccgca gtgtgatggc cgtgcagctr cacagccgcc tgccccgtgg ccgcgtgggc 540
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 600
gaatgtcctc caaggagggtg ataatgtatk tgagcagagg cctgaatgct gaggatagga 660
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 720
gcctctgaca gcttgccctg ctcacttcck cctctgtccc acaggggcgt tcgtggtctg 780
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 840
cccttttagct accttgaact tcagcgggtay gcggcaagca acaaatccac agcccctgat 900
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn 960

```